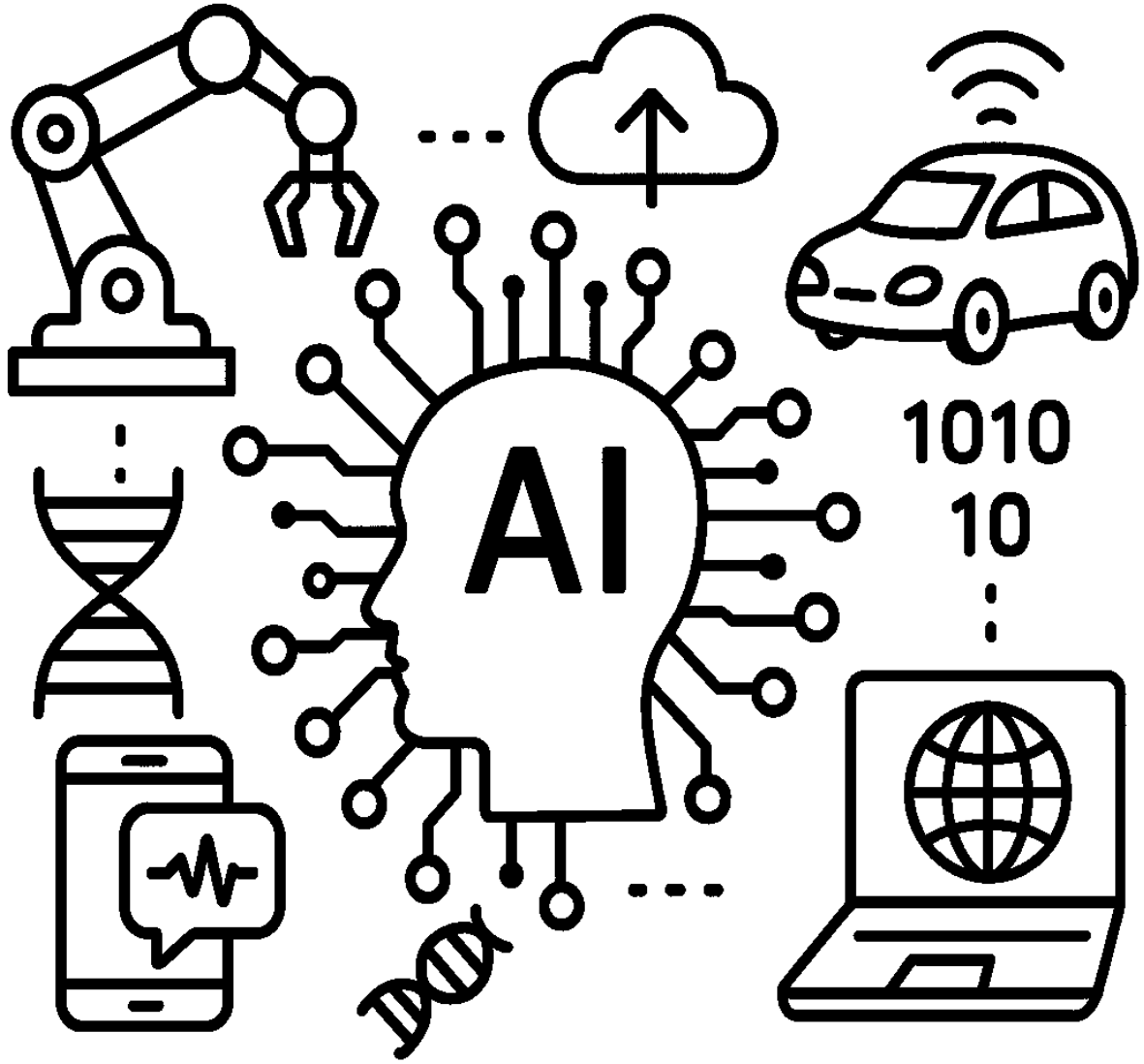
ETICKÉ VÝZVY A MORÁLNE ASPEKTY UMELEJ INTELIGENCIE

Peter Šantavý


AI
1010
10

Fenomén umelej inteligencie ako podstatná súčasť napĺňania zmeny paradigmy...

O S N O V A

- krátke intro do systémov AI -

- problémy, ktorým treba čeliť -

- čo by sme mali, či skôr musíme robiť -

AI – artificial intelligence

O S N O V A

- krátke intro [illegible] AI
- [illegible] treba čeliť
- [illegible]ne mali, či skôr musíme robiť

Každá dostatočne pokročilá technológia je na nerozoznanie od mágie :-)
(Arthur C. Clarke)

AI – artificial intelligence

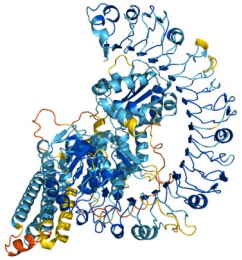
Prečo umelá inteligencia?

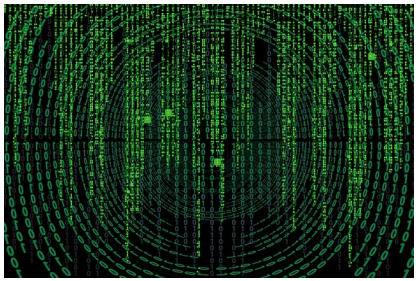
Prečo umelá inteligencia?

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...)

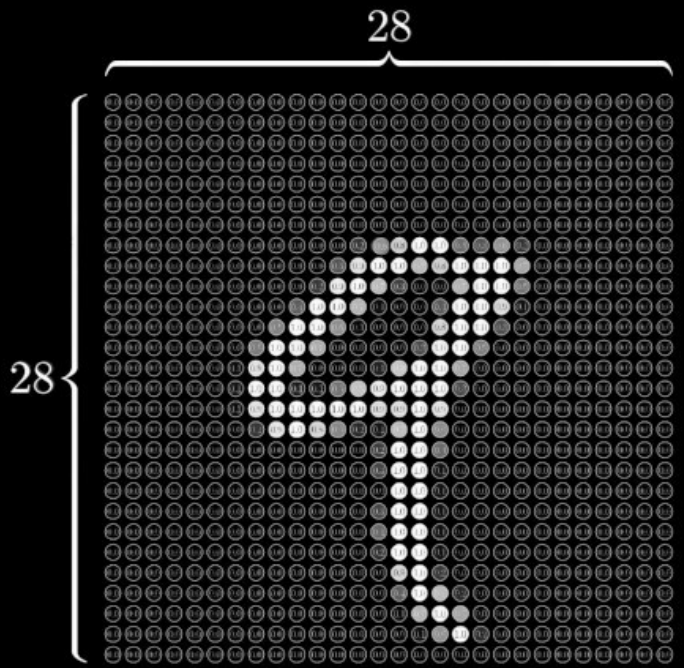




Kredit: 3Blue1Brown

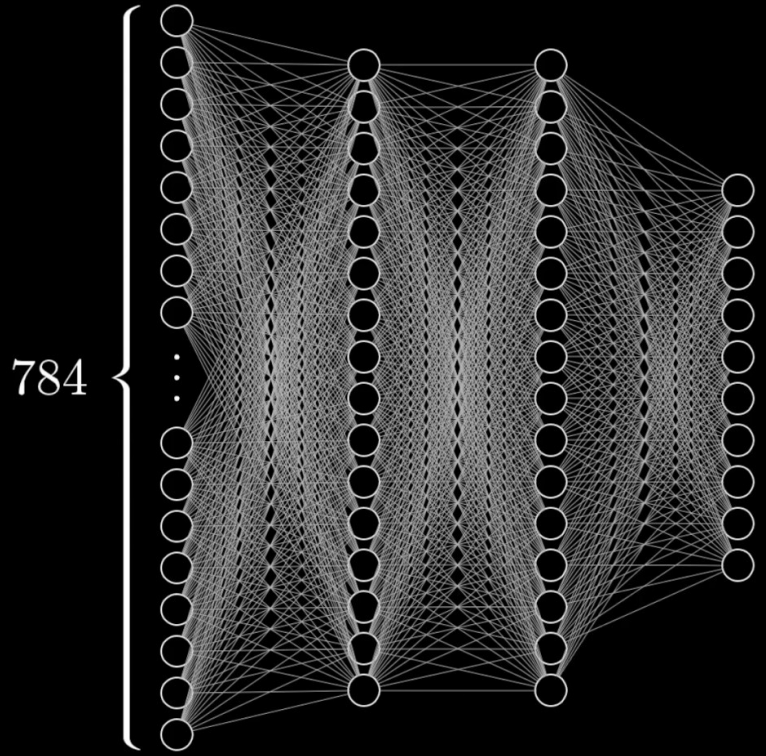
P R E Č O A I A Ž T E R A Z?

Lebo už na to máme schopnosti a využitie...

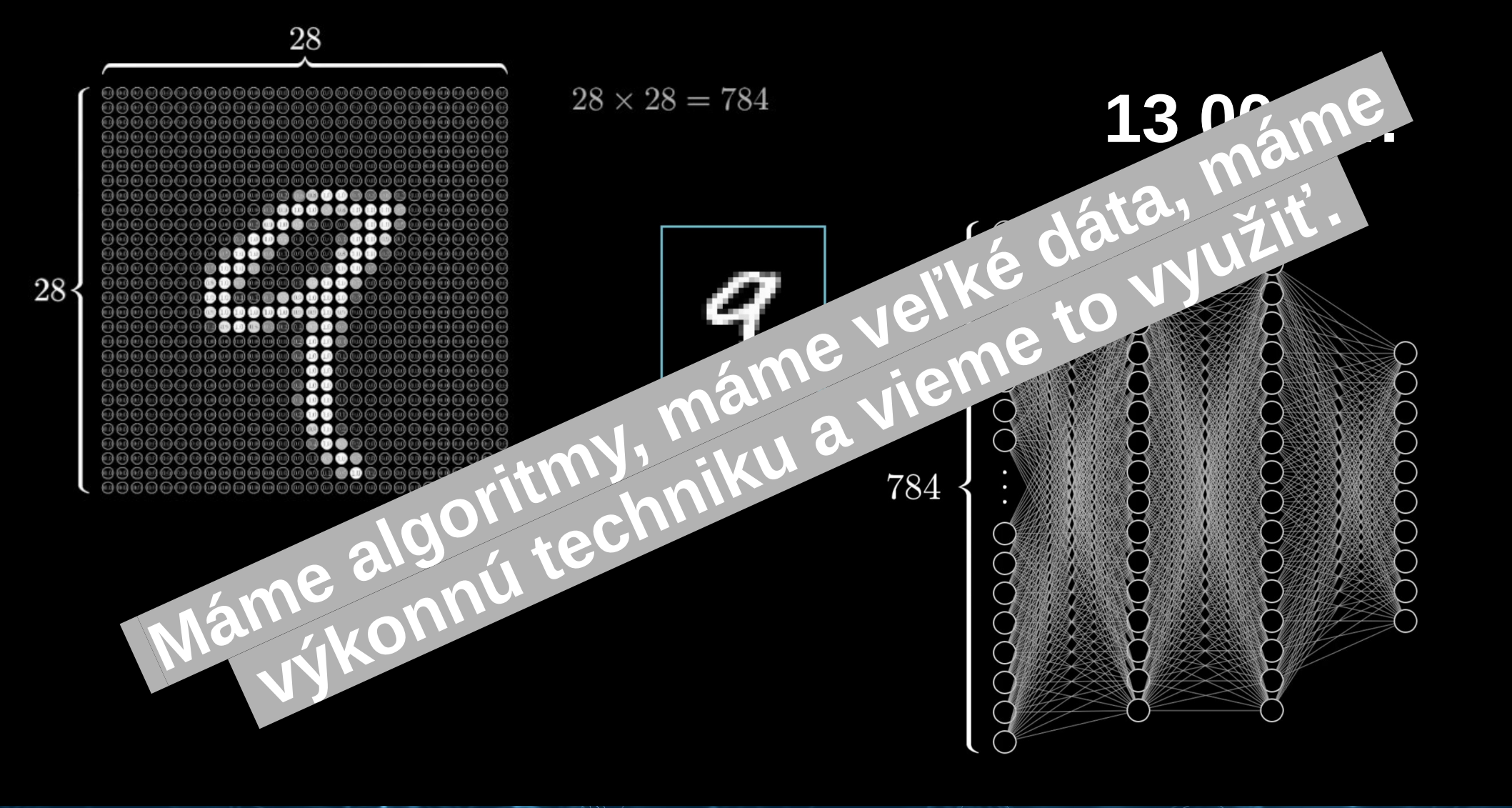


$28 \times 28 = 784$

13 002 !!!

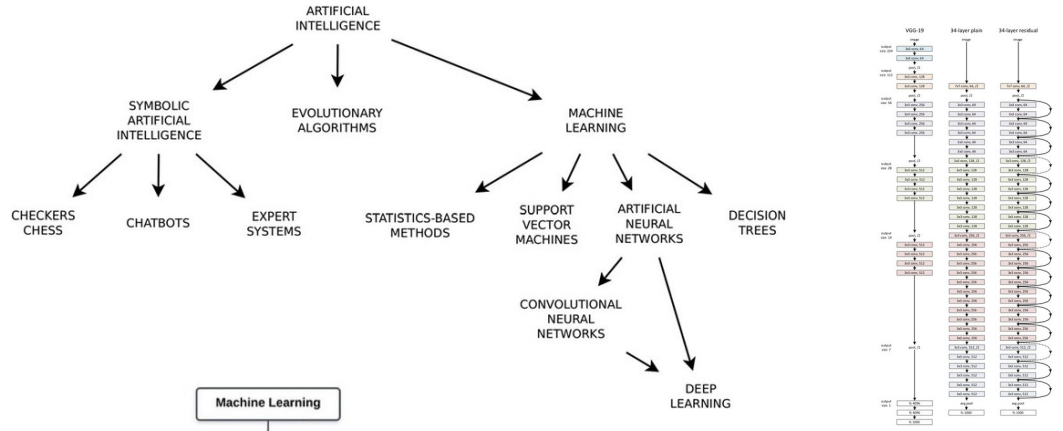


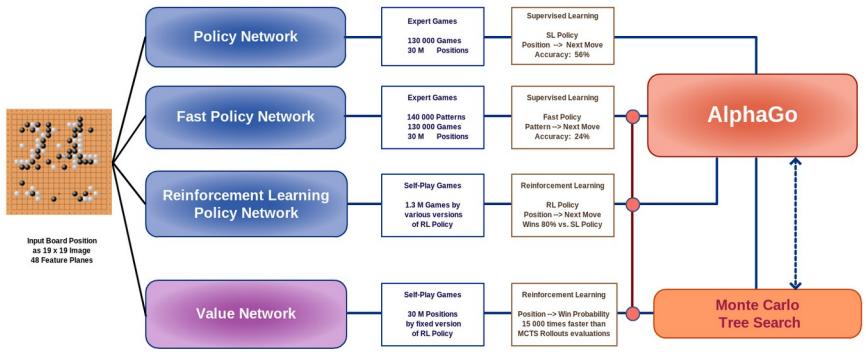
Kredit: 3Blue1Brown

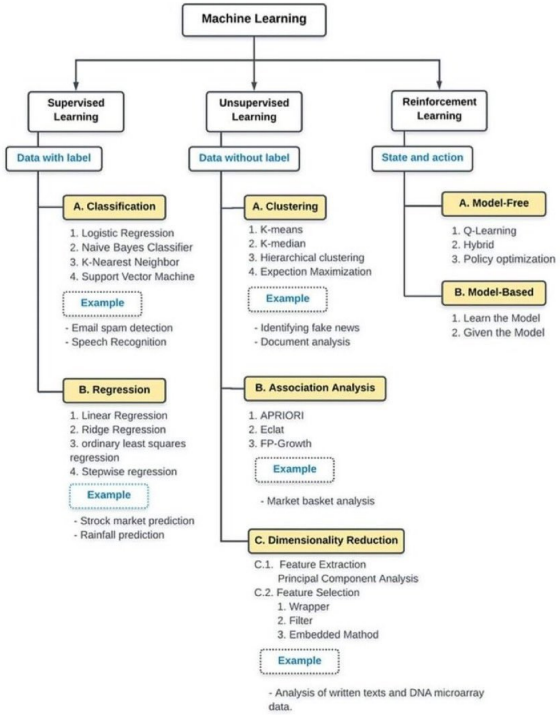


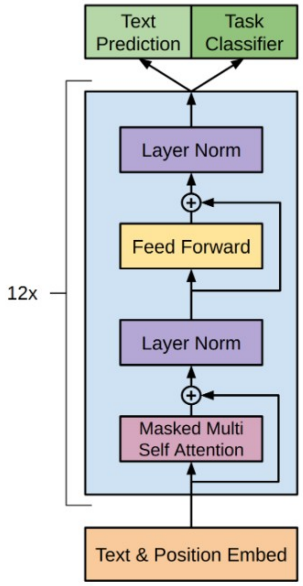
Máme algoritmy, máme veľké dáta, máme výkonnú techniku a vieme to využiť.

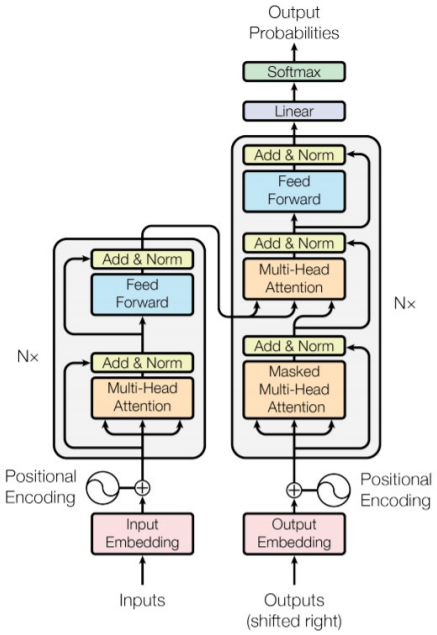
Kredit: 3Blue1Brown

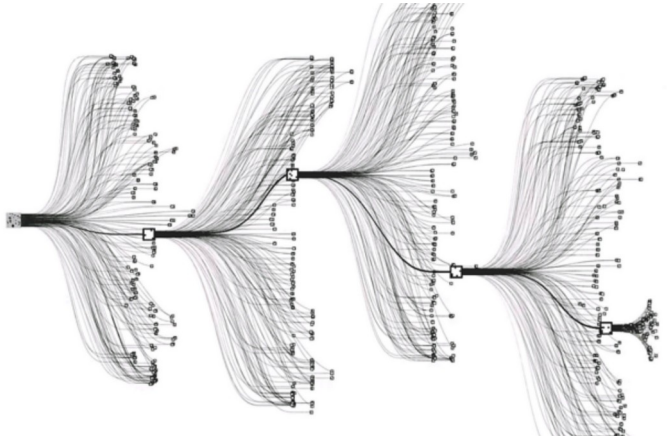












Kredit: Fabio Galbusera, Eric Gaubert, Bob van den Hoek, Seth Weidman, Daniel Dugas

P R E Č O A I?

Existujú ešte nejaké iné dôvody, prečo AI?

...okrem vedeckých a technologických napr. finančné, mocenské, vojenské, ideologické, spoločenské,...

» bytostné «

P R E Č O A I?

Existujú ešte nejaké iné dô[illegible]?

...okrem vede[illegible]logických napr. finančné, moce[illegible], ideologické, spoločenské,...

» bytostné «

Túžba spoznávať, rozvíjať a tvoriť...
...v pozitívnom i negatívnom zmysle...

Definícia umelej inteligencie

Inteligenciu ľudského bytia vnímame ako schopnosť vnímať, chápať a spracovávať informácie, učiť sa, odôvodňovať, plánovať, riešiť komplexné problémy a rozhodovať. Schopnosť tiež používať abstraktné a logické myslenie, predstavivosť a uvažovať o hypotetických možnostiach, jazykový cit, atď. Inteligencia stelesnená a vzťahová...

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

Na prasknutie naplnený kufor rôznych technológií (Marvin Minsky)...

Definícia umelej inteligencie

Inteligenciu ľudského bytia vnímame ako schopnosť vnímať, chápať a spracovávať informácie, učiť sa, odôvodňovať, plánovať, riešiť komplexné problémy a rozhodovať. Schopnosť tiež používať abstraktné a logické myslenie, predstavivosť a uvažovať o hypotetických možnostiach, jazykový cit, atď. Inteligencia stelesnená a vzťahová...

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

Na prasknutie naplnený kufor rôznych technológií (Marvin Minsky)...

Funkcionalistická umelá inteligencia vs. komplexná inteligencia ľudskej bytosti!

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.

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.

Základné delenie – symbolická a subsymbolická AI

Subsymbolická AI – napodobňuje myšlienkové procesy, ktoré by sme mohli nazvať niekedy nevedomými, či automatickými, a ktoré sú základom rýchleho vnímania (fast perception; rozpoznávanie tvárí, identifikácia hovorených slov,...).

Subsymbolická AI – tiež zjednodušene povedané – sa v mnohých prípadoch snaží emulovať činnosť mozgu na úrovni neurónov. Subsymbolické systému sú navrhnuté tak, aby sa učili vykonávať úlohy na základe dát.

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.

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.

Základné delenie – ANI a AGI

Všeobecná umelá inteligencia (AGI) – tzv. silná (strong) a všeobecná (general) umelá inteligencia.

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**.

*Rozlišujeme rôzne stupne AGI: **BAGI** (below-human), HLAGI/HAGI (human-level AGI), MAGI (moderately-superhuman AGI), SAGI/ASI (superintelligent AGI).*

Základné delenie – ANI a AGI

Všeobecná umelá inteligencia (AGI) – t… všeobecná (general) umelá inteligencia.

Všeobecná, lebo **dokáž… intelektuálnu úlohu a má schopnos…** …ať, či adaptovať naučené schopn…

…zumie tomu, čo rieši a vykonáva.

*…jeme rôzne stupne AGI: **BAGI** (below-human), HLAGI/HAGI (human-level AGI), MAGI (moderately-superhuman AGI), SAGI/ASI (superintelligent AGI).*

Nástup AGI: potrebné sú koncepčné prelomy a posun of striktne funkcionalistickej AI k niečomu viac

Generatívne systémy AI (genAI)

Princíp činnosti genAI

– algoritmy genAI sú **schopné podľa vstupnej sekvencie generovať rôznorodý obsah** (text, obrázky, audio, video,...), ktorý napodobňuje ľudský obsah a schopnosť rozumieť

– výstupy sú generované na základe predtrénovaných vzorov a zložitých algoritmov

Technologické aspekty genAI

– data pre-processing -> foundation model -> fine tuning / assistant model -> RLHF

– tokenization, embedding (+ encoders), transformers (+S/MH attention), decoders,...

Mnohé smery súčasného i očakávaného vývoja

– **zvyšovanie kvality** odpovedí a výkonu (škálovanie, algoritmy, destilácia, quantization,...)

– **zlepšovanie “myslenia”** (reasoning modely)

– **prepojenie s reálnym svetom** (agent AI)

– RAG a prispôsobené modely

– multimodálne modely

.....

Generatívne systémy AI (genAI) – prečo?

Evidentný technologický posun

– činnosť podobná ľudskému mysleniu na základe naučených vzorov a technológie transformerov dosahujúca **pozoruhodné výsledky a širokú škálu využitia**

– schopnosť čerpať z extrémneho množstva zdrojov a pracovať s bežnými dátami

– potenciál učiť sa vzory v dátach, kontext aj sémantiku jazyka, porozumieť dlhým konverzáciám, poskytovať súvislé a kontextovo relevantné odpovede

Zmena v spôsobe využívania nástrojov AI

– konverzačné rozhranie umožňujúce komunikáciu v bežnom jazyku formou dialógu

– pokročilá – akoby tímová – **spolupráca medzi človekom a strojom**

– riešenie úloh reálneho sveta prostredníctvom komunikácie s genAI (agenti AI)

Generatívne systémy AI (genAI) – prečo?

Evidentný technologický posun

– činnosť podobná ľudskému mysleniu na základe nau… …ogie transformerov dosahujúca **pozoruhodné výsl… …zitia**

– schopnosť čerpať z extrémneho mn… …s bežnými dátami

– potenciál učiť sa vzory v dát… …a jazyka, porozumieť dlhým konverzáciám, poskyt… …relevantné odpovede

Zmena … …nia nástrojov AI

… umožňujúce komunikáciu v bežnom jazyku formou dialógu

…koby tímová – **spolupráca medzi človekom a strojom**

…nie úloh reálneho sveta prostredníctvom komunikácie s genAI (agenti AI)

Pri správnom používaní sa systémy genAI stávajú silným technologickým prostriedkom!

Moderné systémy AI – dôsledky...

Súčasť paradigmatickej zmeny informačnej a znalostnej spoločnosti

Využívanie nástrojov genAI smeruje k prirodzenému ľudskému spôsobu komunikácie a k radikálnej zmene obsahu i spôsobu práce.

Dôsledky nielen na rozvoj znalostnej spoločnosti, ale – predovšetkým – **na ľudskú psychiku a sociologický rozmer spoločnosti**.

Strategické technológie

Najlepšie a najsofistikovanejšie systémy AI s očakávaným potenciálom rozvoja môžeme považovať za **strategické technológie, ktoré majú potenciál meniť svet – k lepšiemu, ale i k horšiemu**...

O S N O V A

▪ krátke intro do systémov AI ▪

▪ problémy, ktorým treba čeliť ▪

▪ čo by sme mali, či skôr musíme robiť ▪

AI – artificial intelligence

O S N O V A

- krátke intro [illegible] ov AI ▪
- [illegible] treba čeliť ▪
- [illegible] čo by sme mali, či skôr musíme robiť ▪

Nič veľkého nevstúpi do života smrteľníkov bez prekliatia. (Sofokles)

AI – artificial intelligence

S čím sa potýkame?

» **technologické riziká a limity**

» **nesprávne používanie a zneužitie systémov AI**

» **dôsledky pre človeka a spoločnosť**
sociologické, psychologické, civilizačné,...

S čím sa potýkame?

» technologické riziká a li

» nespráv dzivanie s ystémov AI

» dôsledk a spoločnosť

s psychologické, civilizačné,...

AI systémy myslenie len napodobňujú!

AI systémy "rozmýšľajú" diametrálne odlišne, inak než ľudia!

Technologické riziká a limity ANI

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)
- povery (superstition)
- klamanie hlbokých sietí a ich zraniteľnosti (adversarial examples, hacking,...)
- špecifické zlyhania generatívnych systémov (jailbreak: roleplay, base64, UTS, prompt injections, data poisoning / backdoor attacks,...)

Technologické riziká a limity ANI

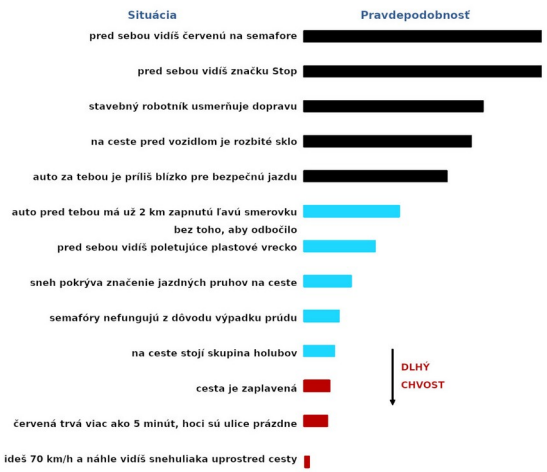
Zraniteľnosti, slabiny a klamanie systémov strojov

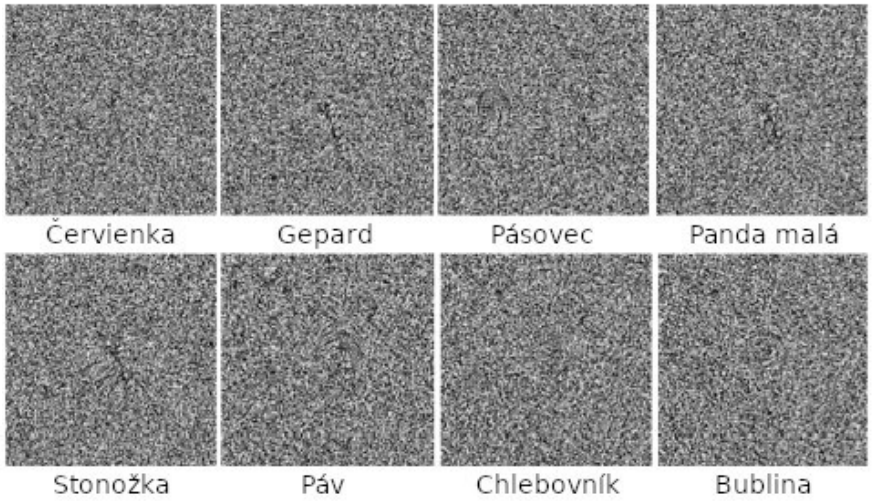
- malá množina trénovacích dát (training data
- nesprávne zvolená/nekvalitná mno
- nadmerné prispôsobovani
- efekt dlhého chvo
- povery (s
- kl
- pecifi

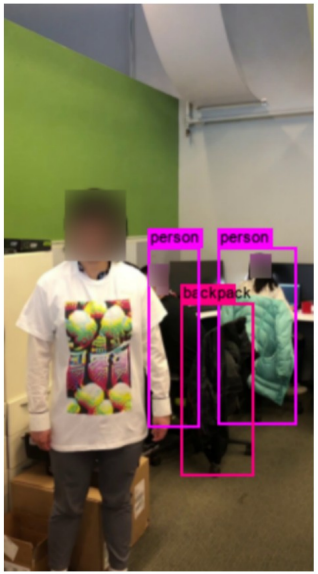
ti a ich zraniteľnosti (adversarial examples, hacking,...)

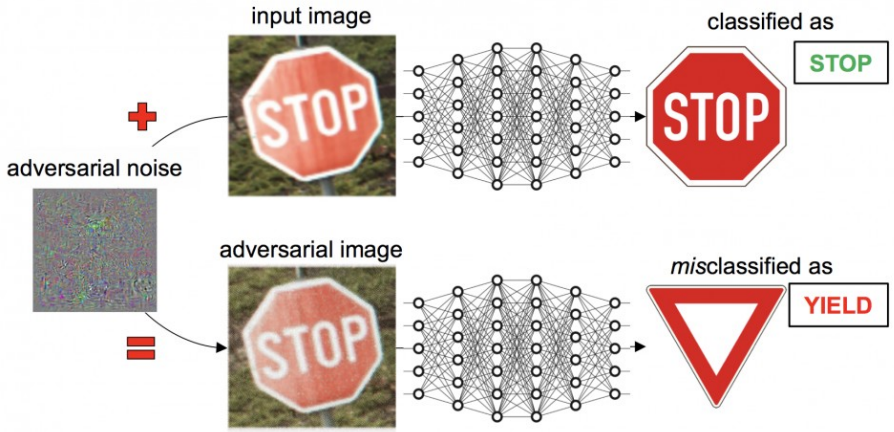
ynania generatívnych systémov (jailbreak: roleplay, base64, UTS, prompt injections, data poisoning / backdoor attacks,...)

Zlyhávajú (nielen) pri nesprávnom použití ...
Ešte viac pri zámernom zneužití...

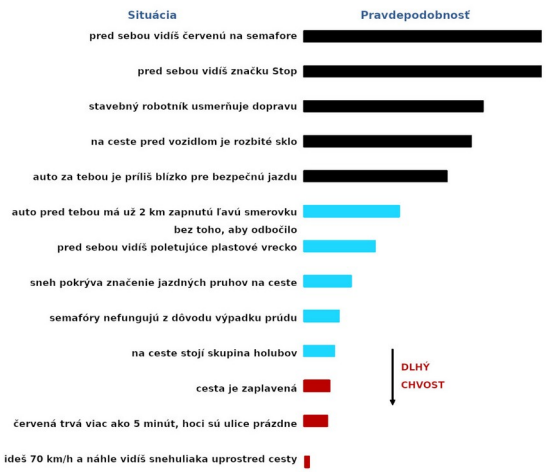


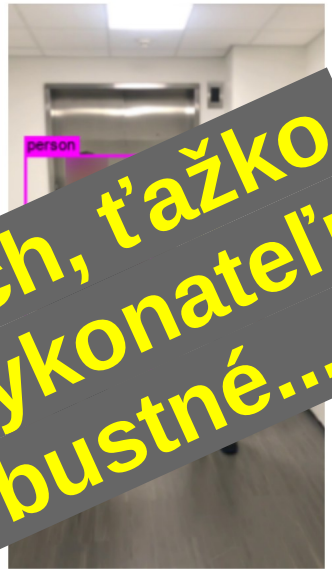


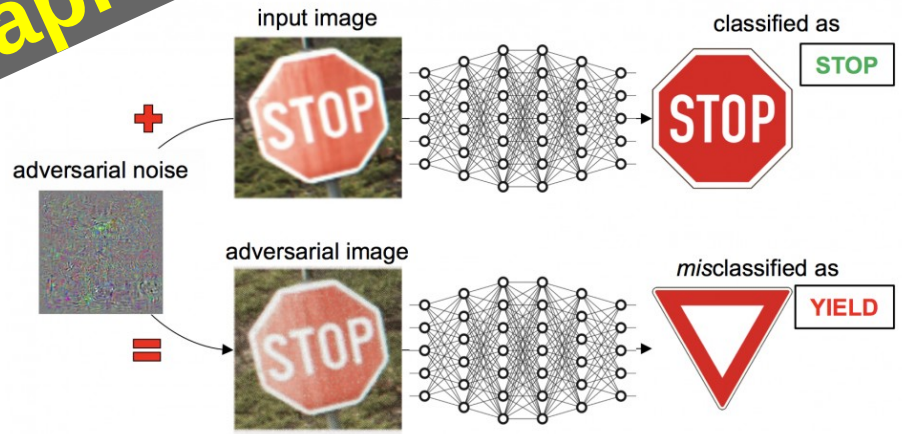




Kredit: Melanie Mitchell – upravené autorom, autor, K. Xu et al., Sourav Agarwal







Zlyhania sú odlišné od ľudských, ťažko predvídateľné, niekedy ľahko vykonateľné a mnohokrát prekvapivo robustné...

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

Procesné riziká a limity ANI

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.

Etické výzvy, doteraz rozoberané ako súčasť návrhu a realizácie systémov AI, sa rozširujú aj o oblasť etiky použitia, resp. riziká zneužitia.

Sekundárne 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 v poslednej dekáde **požadovaný výpočtový výkon i celková réžia zvyšuje exponenciálne**

– **extrémna komplexnosť** je rizikovým faktorom bezpečnosti a stability fungovania systémov

Sekundárne technologické rizikové faktory

Kybernetická bezpečnosť – je to napadnuteľ

– **vektor útoku** na zraniteľnosti AI

– **neexistuje dokonale bezpe** **tém**

Infraštruktú **ie**

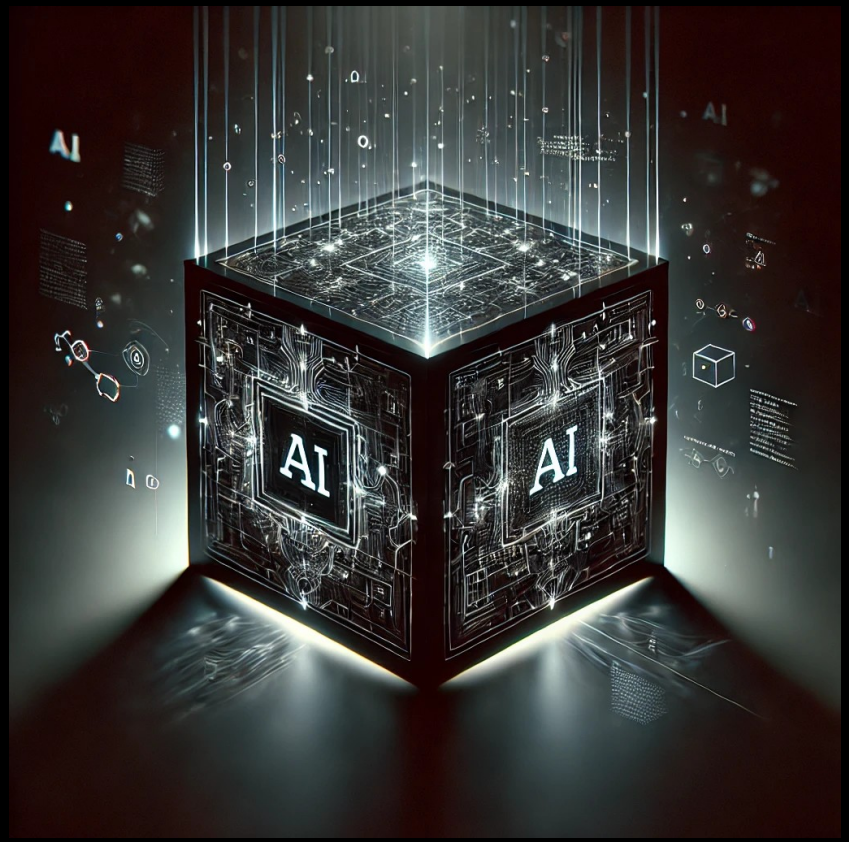
– s r enia sa v poslednej dekáde **požadovaný**

rezia zvyšuje exponenciálne

na komplexnosť je rizikovým faktorom bezpečnosti a stability fungovania systémov

Problémom nie je len zlyhanie AI, ale i človek: zodpovednosť vývoja, etika použitia, riziká zneužitia...

BLACK BOX



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.**

BLACK BOX

Prakticky nevieme, zák e čoho robia hlbok te svoje rozhodnutia.

Pracujeme na vysvetliteľných a interpretovateľných systémoch AI ;-)

Zásade nevieme,
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.

Vybrané dôsledky a riziká pre človeka

Extrémne zhromažďovanie dát

– neustály dohľad a sledovanie bez kontroly

– psychologické profily a modely predikcie nášho správania

– manipulácia a ovplyvňovanie nášho vnímania a konania

– vytváranie sociálnych bublín a rozdelenia, relativizácia hodnôt a pravdy

– technológie AI vedú k závislosti, depresiám, úzkostiam, digitálnej demencii a tzv. mozgovej hnilobe

Vybrané dôsledky a riziká pre človeka

Extrémne zhromažďovanie dát

– neustály dohľad a sledovanie bez k

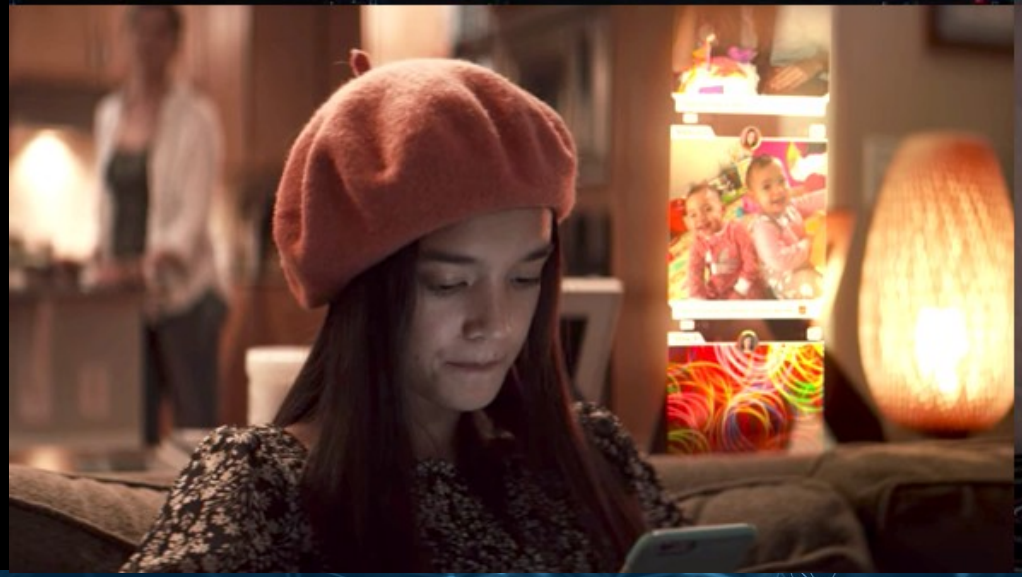
– psychologické profily a sno správania

– manipulácia sno vnímania a konania

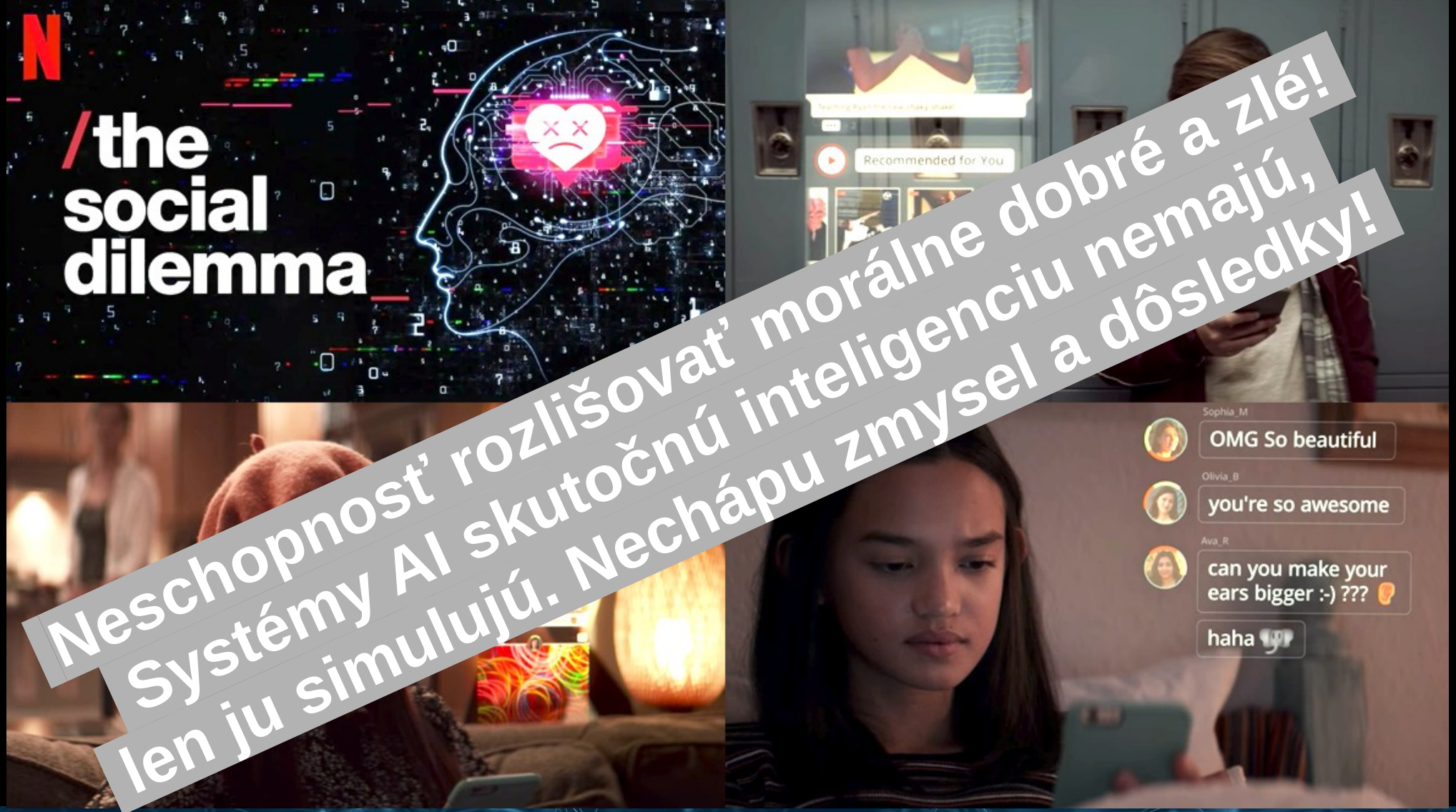
– v soci bublín a rozdelenia, relativizácia hodnôt a pravdy

– technológie AI vedú k závislosti, depresiám, úzkostiam, digitálnej demencii a tzv. mozgovej hnilobe

Ide o riziká, ktoré si mnohokrát ani neuvedomujeme!!!



Kredit: The Social Dilemma



Neschopnosť rozlišovať morálne dobré a zlé! Systémy AI skutočnú inteligenciu nemajú, len ju simulujú. Nechápu zmysel a dôsledky!

Kredit: The Social Dilemma

Vybrané dôsledky a riziká pre spoločnosť

Dohľad a sledovanie ako základ...

– **“z princípu” neustály dohľad a sledovanie, ktoré je ťažko kontrolovať**

Riadenie a fungovanie štátu

– **nezvládnuté a neadekvátne algoritmické riadenie štátu / algokracia**
(súdnictvo, zdravotníctvo, bankový sektor,...)

– **sociálny kreditový systém**

– **neetické vojenské nasadenie**

– **budúcnosť práce a fungovania spoločnosti (mení sa jej paradigma)**

Mocenské, vojenské, ideologické zneužitie AI!

Kredit: Bertelsmann Stiftung

Vybrané riziká generatívnych systémov AI

Halucinovanie

– vymýšľanie si odpovedí, ktoré systém predkladá ako relevantné a správne

Predsudky a neobjektívne výstupy

– riziko poloprávd a nesprávnych odpovedí

Nejasný spôsob narábania s údajmi

– únik dôverných dát, problémy s ochranou osobných údajov a autorskými právami

Vybrané riziká generatívnych systémov

Halucinovanie

– vymýšľanie si odpovedí, ktoré s… tné a správne

Predsud…

…ia s údajmi

– únik dôverných dát, problémy s ochranou osobných údajov a autorskými právami

Generatívne nástroje AI nemôžeme vnímať ako faktografické, spoľahlivé a etické zdroje! Potrebujú človeka, ktorý ich vie správne používať…

Etické problémy (nielen) generatívnych systémov

Neschopnosť rozlišovať morálne dobré a zlé – systémy AI skutočnú inteligenciu nemajú, len ju simulujú. **Nechápu zmysel a dôsledky**.

Riziko digitálnej demencie – prichádza k degradácii intelektuálnych schopností a k dopamínovej závislosti na základe ponorenia sa do virtuálneho sveta, v ktorom systémy umelej inteligencie v čoraz väčšej miere **suplujú** kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení.

Digitálne rozdelenie – riziko digitálneho rozdelenia (digital divide) i v oblasti umelej inteligencie (tí, ktorí nepoužívajú / sú manipulovaní / využívajú AI).

RIZIKO DIGITÁLNEJ DEMENCIE

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

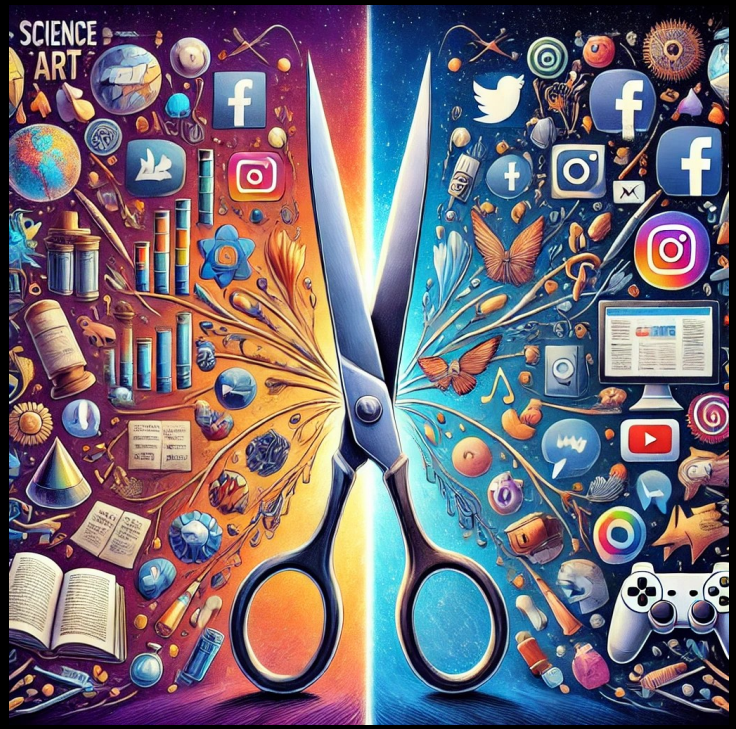
Môže prichádzať k degradácii intelektuálnych schopností a k dopamínovej závislosti na základe ponorenia sa do virtuálneho sveta, v ktorom systémy AI v čoraz väčšej miere suplujú kognitívne činnosti človeka, a to spôsobom, na ktorý nie sme evolučne vôbec pripravení. Mozog sa mení...

RIZIKO DIGITÁLNEJ DEMEN

Dlhodobé a nesprávne využíva , osobitne generatívnych A ciálnych sietí, prináša inten se kognitívne ychniku.

k degradácii intelektuálnych a k dopamínovej závislosti na základe ma 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...

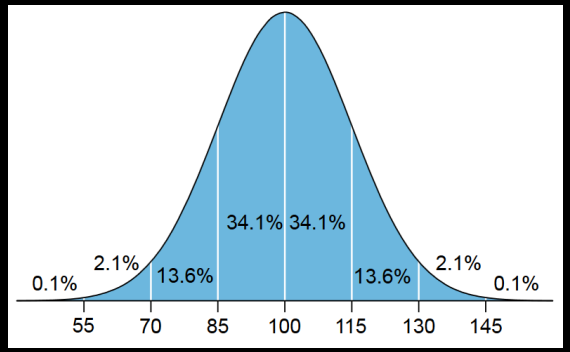
Vďaka systémom AI môžeme postupne strácať časť našich schopností a psychickej odolnosti!



INTELIGENČNÉ NOŽNICE, KTORÉ SA OTVÁRAJÚ

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

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



Kredit: wikipedia.org

Etické problémy (nielen) generatívnych systémov

Mozgová hniloba (brainrot) – postupná strata schopnosti sústrediť sa a vnímať hlbší kontext. Vytvárajú sa dopamínové dráhy závislosti a návyk na také hodnoty adrenalínu a dopamínu, na ktoré nie sme evolučne pripravení.

Deepfake – vytváranie dôveryhodných virtuálnych identít ľudí a podvrhnutie ich falošnej komunikácie a konania (klamanie, nekalá konkurencia, pornografia, kriminalita, propaganda,...). Stieranie rozdielov medzi pravdou a lžou.

Strata zmyslu pre realitu – únik do virtuálneho sveta a do vzťahov s AI. Neschopnosť fungovať v reálnom svete a rozvíjať skutočné vzťahy, hodnotové posuny, závislosti, zlyhávanie v reálnom živite (kritické myslenie, sexualita,...).

Etické problémy (nielen) generatívnych ov

Mozgová hniloba (brainrot) – postupná str a vnímať hlbší kontext. Vytvárajú sa dopamín noty adrenalínu a dopamínu, na

Deepfake ľudí a podvrhnutie ich fal konkurencia, pornografia, zdielov medzi pravdou a lžou.

Virtuálne zážitky v mozgu mnohých evokujú tak vysoké hladiny adrenalínu a dopamínu, aké nie sú schopní v reálnom živote dosiahnuť. Únik do virtuálneho sveta...

Stratu ...ntu – únik do virtuálneho sveta a do vzťahov s AI. Nes... fungovať v reálnom svete a rozvíjať skutočné vzťahy, hodnotové po..., závislosti, zlyhávanie v reálnom živite (kritické myslenie, sexualita,...).

Kredit: AP News

Deepfake: klamanie, nekalá konkurencia, pornografia, kriminalita...

Kredit: AP News

Nekonečné možnosti tvorby audiovizuálneho obsahu, ktorý vytvára virtuálny svet a neskutočnú realitu...


Nekonečné možnosti tvorby audiovizuálneho obsahu, ktorý vytvára virtuálny svet a neskutočnú realitu...
Strata zmyslu pre realitu a skutočné hodnoty!


Nekonečné možnosti tvorby audiovizuálneho obsahu, ktorý vytvára virtuálny s
a neskutočnú realitu
Strata zmyslu pr
skutočné hodnoty!
Matrix: Neo bojuje za niečo, čo je skutočné...
“A koho to zaujíma, či je to skutočné?”

Ďalšie etické problémy generatívnych systémov

Neschopnosť robiť vlastné zodpovedné rozhodnutia – delegujeme to na stroje.

Narušenie psychického vývoja priskorým, resp. nezrelým používaním digitálnych a AI technológií (od synaptickej plasticity až po napr. tzv. *virtuálny autizmus...*)**.**

Riziko erózie identity – preceňovanie AI a "splošťovanie" ľudskej komplexnej inteligencie, racionality a celkovej ľudskosti na úroveň technickej inteligencie.

Strata kritického myslenia a nekritické preberanie výsledkov systémov AI spojené s relativizáciou pravdy, vytváraním sociálnych bublín a pod.
Schopnosť a možnosť manipulácie vlastníkmi systémov AI.

Propaganda a manipulácia, nástroj moci a ideológie, algoritmické modelovanie histórie,...

Ďalšie etické problémy generatívnych systém

Neschopnosť robiť vlastné zodpovedné rozhodnutia – dele

Narušenie psychického vývoja priskorým, resp **digitálnych a AI technológií** (od synaptickej plasticity *autizmus...*).

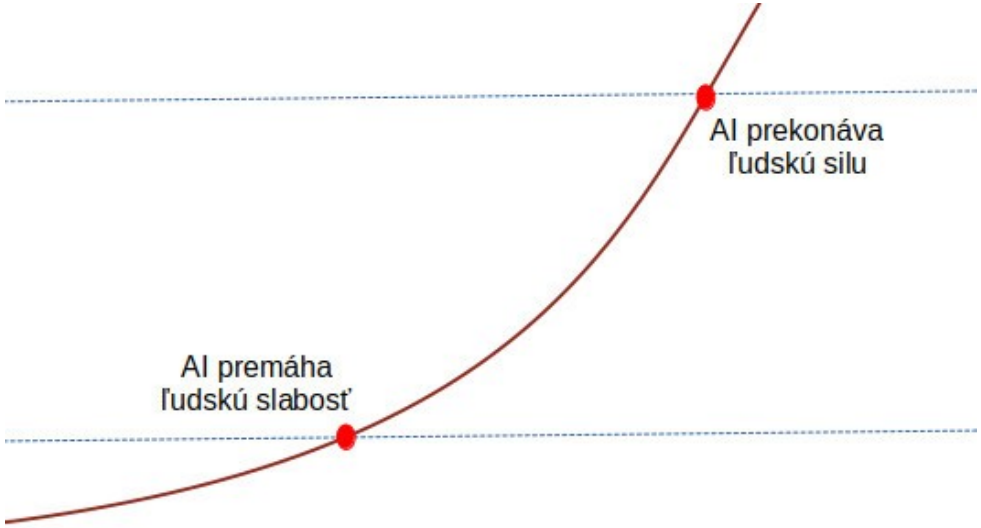
Riziko erózie identity – prec **vanie" ľudskej komplexnej inteligencie, racion** **ti na úroveň technickej inteligencie.**

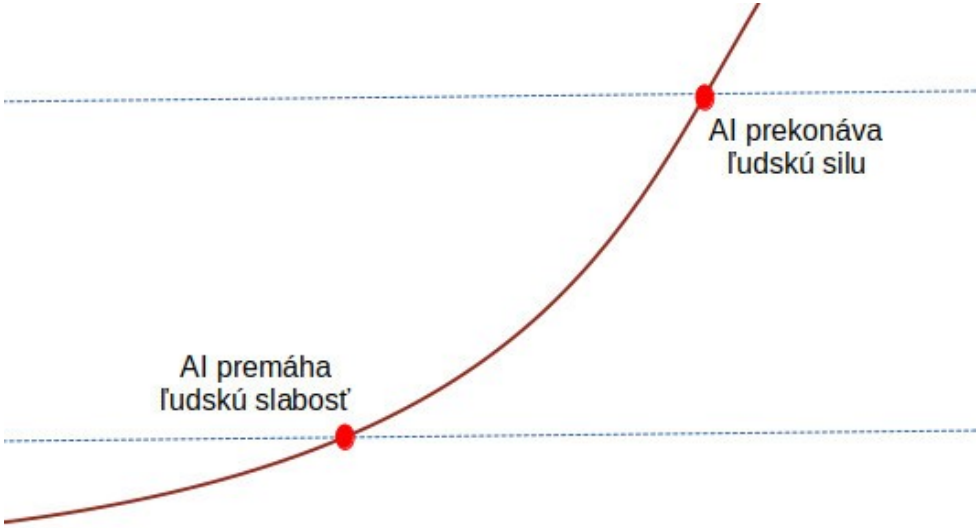
Strata kr **kritické preberanie výsledkov systémov AI spojené** **, vytváraním sociálnych bublín a pod.**

ožnosť manipulácie vlastníkmi systémov AI.

paganda a manipulácia, nástroj moci a ideológie, algoritmické modelovanie histórie,...

Vďaka systémom AI môžeme postupne strácať časť našich schopností a psychickej odolnosti.


AI prekonáva
ľudskú silu
AI premáha
ľudskú slabosť



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.

O S N O V A

▪ krátke intro do systémov AI ▪

▪ problémy, ktorým treba čeliť ▪

▪ čo by sme mali, či skôr musíme robiť ▪

AI – artificial intelligence

Základné postoje vo svetle Zjavenia I.

Človek stvorený na Boží obraz (צֶלֶם אֱלֹהִים, imago Dei)

– **inteligencia a jej používanie** sú jedným zo základných aspektov napĺňania "obrazu"

Užívanie (podmanenie) zvereného sveta je v Gn 1,28 vyjadrené hebrejskými výrazmi „podmaňte si ju“ (וְכִבְשֻׁהָ) a „panujte“ (וּרְדוּ)

– užívať, rozvíjať, chrániť a z generácie na generáciu si odovzdávať stvorený svet

– dar inteligencie by sa mal prejavovať **zodpovedným využívaním** rozumu a akýchkoľvek civilizačných vymožeností i technicko-vedeckých schopností pri správe stvoreného sveta

Konfrontácia *chokmá* vs. *sofia* v ketubím (múdroslovnej literatúre SZ)

– ovocím tejto konfrontácie nie je zavrhnutie civilizačného rozvoja, ale neustále **pozvanie rozvíjať a užívať svet podľa Božích zákonov, vo vzťahu voči Bohu a pre dobro ľudí**

Základné postoje vo svetle Zjavenia II.

„Choďte teda, učte všetky národy...“ (Mt 28, 19) a „Choďte do celého sveta a hlásajte evanjelium všetkému stvoreniu“ (Mk 16, 15)

– nadčasové a všeobímajúce poslanie, zahŕňajúce aj **civilizačný rozvoj**

Sv. Pavol na misijných cestách

– **v aténskom Areopágu** (2. misijná cesta) ohlasuje jazykom helénskych mysliteľov

– v pastoračnom pôsobení v helénskom svete (jazykom synkretizmu Kolosanov,...)

– vyučovanie v škole gréckeho filozofa a učiteľa rétoriky Tyranna v Efeze (3. MC)

Nasledovníci Pavla v civilizačnej konfrontácii

– filozof a mučeník sv. Justín, predstavitelia alexandrijskej a antiochijskej školy,...

– scholastici, scholastická metóda, Occamova britva a následný **vedecký rozvoj**

Základné postoje v učení Magistéria

Dokumenty pápežov – od komunikačných prostriedkov k IKT a AI

– od Gregora XVI. (1831-1846) až po pápeža Františka

– od IKT k virtuálnemu svetu a AI v dokumentoch od Pia XI. až po Sv. otca Františka

– prorocké pozvanie a poslanie v encyklike ***Redemptoris missio*** sv. Jána Pavla II.

Koncilové dokumenty

– dekrét o spoločenských komunikačných prostriedkoch ***Inter Mirifica***

– pastorálna konštitúcia o Cirkvi v dnešnom svete ***Gaudium et spes***

– celkovo 9 koncilových dokumentov v nejakej miere sa zaoberajúcich aj SKP a IKT

Postoje a dokumenty Cirkvi na Slovensku

– pastoračné plány KBS a rady KBS

– dokumenty biskupov a diecéznych synôd

Magistérium a problematika AI

Rome Call for AI Ethics (2020 – súčasnosť)

– **etické výzvy a princípy** v prístupe, vývoji, realizácii, prevádzke i využívaní systémov AI

– snaha o zjednocujúce pôsobenie a angažovanie náboženstiev, štátov, korporácií,...

– konferencia, dokument a participácia aktérov, ktorých zoznam sa rozširuje i v súčasnosti

Antiqua et nova (2025)

– komplexný dokument o vzťahu medzi ľudskou a umelou inteligenciou

– praktický a odborný dokument: pohľad odborníkov z rôznych uhľov pohľadu a profesií

– "**technologický pokrok je súčasťou Božieho plánu pre stvorenie**"

Dokumenty riadneho Magistéria o AI

– príhovory a dokumenty pápeža Františka

– dokumenty vatikánskych dikastérií

Na dobro človeka zameraná umelá inteligencia

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**, resp. klasickej filozofickej antropológie (predovšetkým biologickej a kultúrnej)

– zachovávať **ľudskú dôstojnosť** a podporovať integrálny rozvoj ľudskej osoby a 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 stvorený svet

Na dobro človeka zameraná umelá inteligencia =

= dôveryhodné systémy AI, ktoré musia byť

funkčné a užitočné – navrhnuté a realizované tak, aby vykonávali požadovanú činnosť

legálne – musia vyhovovať požadovaným normám, zákonom i reguláciám a spĺňať všetky platné zákony a predpisy

etické – musia rešpektovať etické zásady a hodnoty

odolné, resp. robustné* – dosahujúce potrebné štandardy bezpečnosti a spoľahlivosti nielen z technologického hľadiska, ale zohľadňujúce aj sociálne prostredie a dopady na spoločnosť

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

Nutné podmienky seriózneho využívania AI

Human-centred AI a trustworthy AI – **nutné princípy** akéhokoľvek výskumu, vývoja, realizácie alebo nasadenia systémov AI

Funkčnosť, etika, legálnosť a odolnosť – **nutná podmienka** vývoja, nasadenia, prevádzky a využívania systémov AI

Vzdelávanie, osveta i prevencia a formácia morálnych postojov tvorcov, 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 umelej inteligencie

Čo treba robiť na spoločenskej úrovni

Mať jasne definované etické princípy

– ich chápanie, hodnotový systém, spoločenský konsenzus

Primerané a široko akceptované regulácie

– 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

– adekvátne jednotlivým profesiám a spôsobu využívania systémov AI

Jasný postoj a poznanie (osobná rovina)

Jasný postoj k technológiám (kriticky pozitívny a postavený na hodnotách, pre veriaceho i na Zjavení)

Základné poznanie moderných technológií a systémov AI so schopnosťou **vedieť využiť ich potenciál**

Poznanie a chápanie rizík, ktoré sú spojené s AI a modernými technológiami a ich dôsledkov

– vnímanie mnohých rizikových faktorov je viazané aj na náš hodnotový systém a svetonázor

Spôsob využívania systémov AI (osobná rovina)

V kontexte čností používať na dobré *ciele*
– **múdrosť, pravda a spravodlivosť**, na ktorých je postavená autentická schopnosť tvoriť, spravovať a rozvíjať...
– **miernosť** pre primerané a **mravnosť** pre osožné používanie AI

Používať nástroje AI s pokorou vo vedomí vlastných mantinelov a slabostí
– bez pravdy nielen o technológiách moderného sveta, a predovšetkým **pravdy o sebe** to nepôjde (moje obmedzenia, slabosti a zranenia...)

V askéze
– schopnosť **byť off-line** mimo systémov AI a virtuálneho sveta
– **vytrvať v zásadách** správneho využívania
– vedome **realizovať činnosť, ktorá je "nudná"** a nie je spojená s dopamínovým a adrenalínovým efektom (dopamínový pôst)
– **činnosť v reálnom svete** – fyzická, kreatívna, služba, vzťahy

Spôsob využívania systémov AI (osobná rov

V kontexte čností používať na dobré *ciele*

– **múdrosť, pravda a spravodlivosť**, na ktorých je
schopnosť tvoriť, spravovať a rozvíjať...
– **miernosť** pre primerané a **mravnos** nie AI

Používať nástroje AI s pok **stných mantinelov a slabostí**

– bez pravdy nielen o ného sveta, a predovšetkým **pravdy o sebe** to nepô , slabosti a zranenia...)

V as

-line mimo systémov AI a virtuálneho sveta
ásadách správneho využívania
nie **realizovať činnosť, ktorá je "nudná"** a nie je spojená s dopamínovým a adrenalínovým efektom (dopamínový pôst)
– **činnosť v reálnom svete** – fyzická, kreatívna, služba, vzťahy

V kontexte stanovených cieľov nebáť sa byť kreatívny, tvorivý a zodpovedne odvážny.

V službe iným (osobná rovina)

Formovať tých, ktorí sú mi zverení

– **chrániť deti** pred prístupom k AI, jej obsahu a nesprávnemu využívaniu
– **formovať a vychovávať** k správnemu používaniu
– **prakticky viesť** k správnemu využívaniu iných
(otcova dielňa, pedagóg, výskumný tím...)

Angažovať sa v oblasti AI

– etické a transparentné systémy AI v ochrane ľudskej dôstojnosti, spoločného dobra, sociálnej spravodlivosti, pravej slobody a práva

AI vo vzdelávaní

AI ako neriadená strela

– **mnohoraké využívanie (i zneužívanie)** nástrojov AI študentami

– **otvorený príbeh** s rôznymi možnosťami využívania a ovocím, čo to prinesie

AI vo vzdelávaní

AI ako neriadená strela

– **mnohoraké využívanie (i zneužívanie)** nástrojov AI študentami
– **otvorený príbeh** s rôznymi možnosťami využívania a ovocím, čo to prinesie

Príklad pravidiel využívania nástrojov AI na UK (smernica + komisia)

– **UK podporuje aktívne využívanie AI** a získavanie potrebných zručností
– **cieľ: tvorivé, bezpečné, zodpovedné, etické a transparentné používanie**
– nástroje AI majú byť **podporou a pomocou pri štúdiu**
– **priame vypracovanie celých textov** pomocou nástrojov AI je **plagiarizmus**
– používanie nástrojov AI musí byť **riadne citované**
– pravidlá stanovujú **vyučujúci** (+ výzva prečo pedagóg a nie AI tútor)
– **vízia:** personalizácia vzdelávania; hĺbková analýza dát a pohľad na štúdium

Smernica rektora UK | Dodatok č.1 (citovanie) | Dodatok č. 2 (využívanie)
Nástroje a využívanie AI na RKCMBF UK

Čo by som chcel záverom akcentovať

Potenciál fenoménu umelej inteligencie

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

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

– **riziká a limity fungovania i problematiku zneužitia**
– **psychologické a spoločenské dopady**

Odvahu, snahu a schopnosť mať “opraty pevne v rukách”

– **etické princípy a regulácie**
– osobný postoj a **spôsob vývoja, nasadenia, prevádzky a využívania**
– **vzdelávanie a osveta**

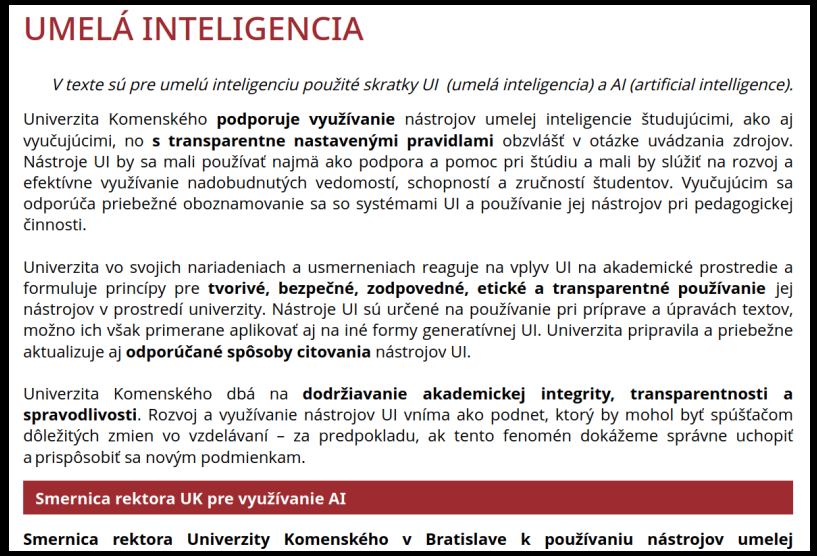
Je na nás, či umelá inteligencia

(a technológie vo všeobecnosti)

bude dobrý sluha alebo zlý pán...

Ďakujem za pozornosť

ThLic. Ing. Peter Šantavý, PhD.
peter@santavy.sk



UMELÁ INTELIGENCIA

V texte sú pre umelú inteligenciu použité skratky UI (umelá inteligencia) a AI (artificial intelligence).

Univerzita Komenského **podporuje využívanie** nástrojov umelej inteligencie študujúcimi, ako aj vyučujúcimi, no **s transparentne nastavenými pravidlami** obzvlášť v otázke uvádzania zdrojov. Nástroje UI by sa mali používať najmä ako podpora a pomoc pri štúdiu a mali by slúžiť na rozvoj a efektívne využívanie nadobudnutých vedomostí, schopností a zručností študentov. Vyučujúcim sa odporúča priebežné oboznamovanie sa so systémami UI a používanie jej nástrojov pri pedagogickej činnosti.

Univerzita vo svojich nariadeniach a usmerneniach reaguje na vplyv UI na akademické prostredie a formuluje princípy pre **tvorivé, bezpečné, zodpovedné, etické a transparentné používanie** jej nástrojov v prostredí univerzity. Nástroje UI sú určené na používanie pri príprave a úpravách textov, možno ich však primerane aplikovať aj na iné formy generatívnej UI. Univerzita pripravila a priebežne aktualizuje aj **odporúčané spôsoby citovania** nástrojov UI.

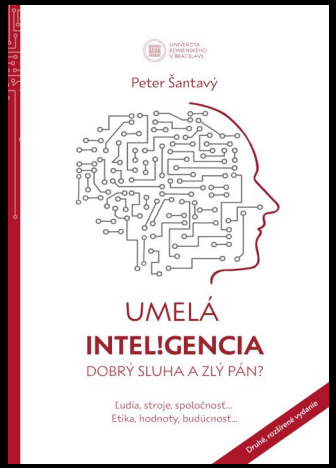
Univerzita Komenského dbá na **dodržiavanie akademickej integrity, transparentnosti a spravodlivosti**. Rozvoj a využívanie nástrojov UI vníma ako podnet, ktorý by mohol byť spúšťačom dôležitých zmien vo vzdelávaní – za predpokladu, ak tento fenomén dokážeme správne uchopiť a prispôsobiť sa novým podmienkam.

Smernica rektora UK pre využívanie AI

Smernica rektora Univerzity Komenského v Bratislave k používaniu nástrojov umelej

Nástroje a informácie k AI v rámci vzdelávacieho procesu na UK

RKCMBF UK - Umelá inteligencia



Niečo na čítanie...

Systémy umelej inteligencie skutočnú inteligenciu nemajú, len ju simulujú

Rozhovor pre NM

Moje knihy z oblasti AI

Etické výzvy a morálne aspekty AI | Anotácia

Fenomén umelej inteligencie už dávno prekročil rámec atraktívnej technológie úzko využiteľnej v špecifických akademických, vojenských či priemyselných nasadeniach. Povedané rečou moderného technokrata – ide o emergentnú a disruptívnu technológiu, teda novú a rozvíjajúcu sa technológiu, ktorá nahrádza a ruší technológie predošlé so schopnosťou zásadne meniť spôsob, akým spoločnosť funguje, pričom má potenciál posunúť civilizáciu ďalej. Čo však už technokrat nepovie – ide o technológiu, ktorú ak nezvládneme, má potenciál nás priviesť k skutočne veľkým problémom.

Ruka v ruke s nárastom potenciálu a dopadu technológií umelej inteligencie sa etika, legálnosť, bezpečnosť i odolnosť týchto systémov a ich nasadenia stávajú alfou a omegou ich zodpovedného využívania pre dobro človeka a spoločnosti.

Príspevok chce poukázať na niektoré riziká a výzvy v technologickej oblasti a ich sociologické, psychologické i civilizačné dôsledky s akcentom na základný etický rámec dôveryhodných systémov umelej inteligencie a ich zodpovedného využívania.