

Dobrý sluha a zlý pán 21. storočia

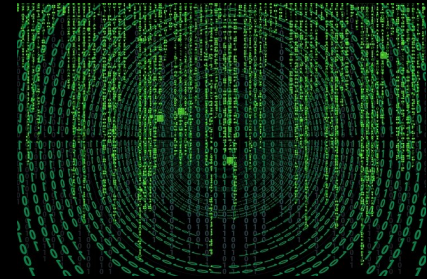
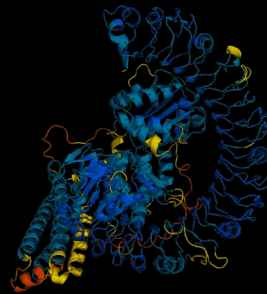
Peter Šantavý

AI – etické aspekty a potencionálne riziká

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...)

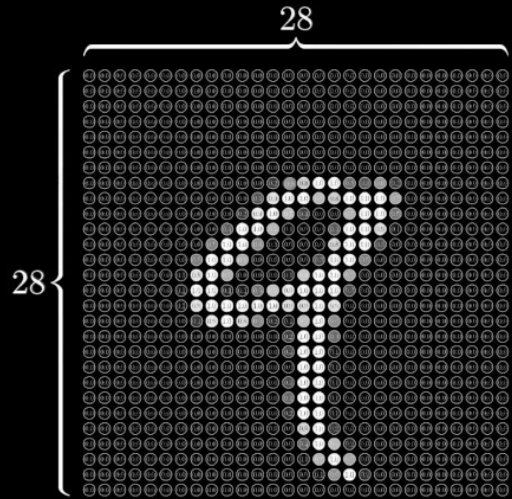


Ked' to algoritmicky nevieme riesit'...

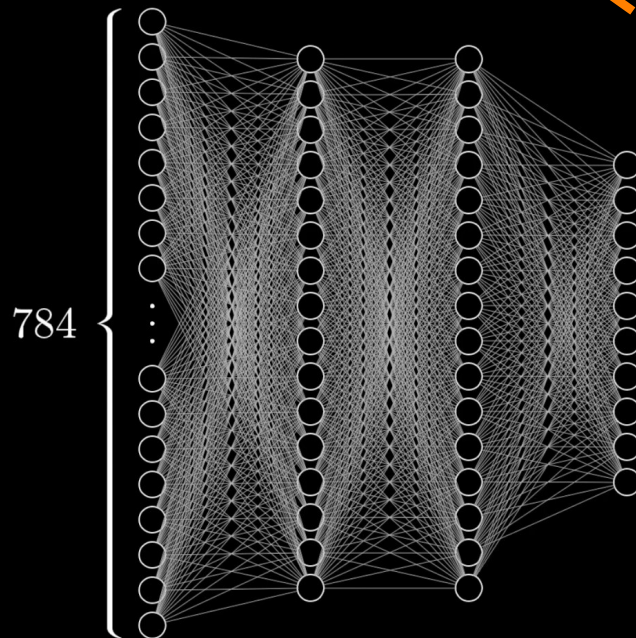
(8, 9) (0, 0) (2, 2) (1, 1) (7, 7) (8, 8) (3, 3) (3, 3)
(0, 0) (4, 4) (6, 6) (7, 7) (4, 4) (6, 6) (1, 1) (0, 0)
(7, 7) (8, 8) (6, 6) (1, 1) (5, 5) (7, 7) (9, 9) (7, 7)
(1, 1) (1, 1) (4, 4) (3, 3) (2, 2) (2, 2) (9, 9) (3, 3)
(1, 1) (1, 1) (0, 0) (4, 4) (6, 6) (0, 0) (0, 0) (6, 6)
(1, 1) (0, 0) (2, 2) (9, 9) (1, 1) (8, 8) (8, 8) (4, 4)
(9, 9) (6, 6) (3, 3) (4, 4) (5, 5) (4, 4) (1, 1) (8, 8)
(3, 3) (8, 8) (5, 5) (4, 4) (7, 7) (7, 7) (4, 4) (2, 2)
(8, 8) (5, 5) (8, 8) (1, 1) (9, 7) (3, 3) (4, 4) (6, 6)
(1, 1) (9, 9) (9, 9) (6, 6) (0, 0) (1, 1) (1, 1) (2, 2)

Ak je to t'azke naprogramovat', nech sa to stroje naučia!

Prečo veľký rozvoj AI až teraz?



$$28 \times 28 = 784$$



13 002 !!!

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

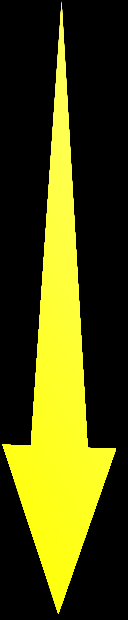
Uvedenie do problematiky AI

- autonómne a adaptívne systémy
- slabá a úzka umelá inteligencia (ANI vs. AGI)
- symbolické a subsymbolické systémy
- generatívne systémy AI



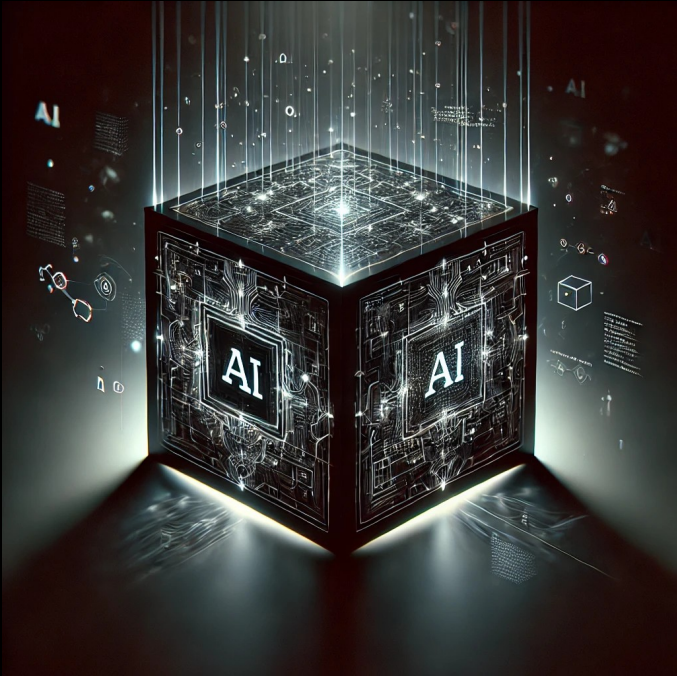
Na môj pomerne rozsiahly popis systémov AI odkazuje zobrazený QR kód...

Na čo treba pamätať

- 
- 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)



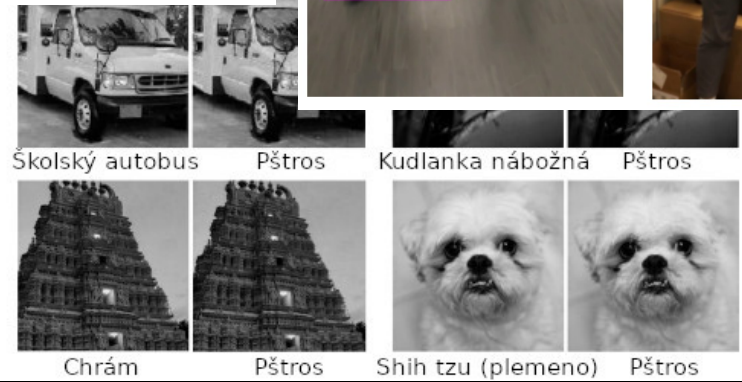
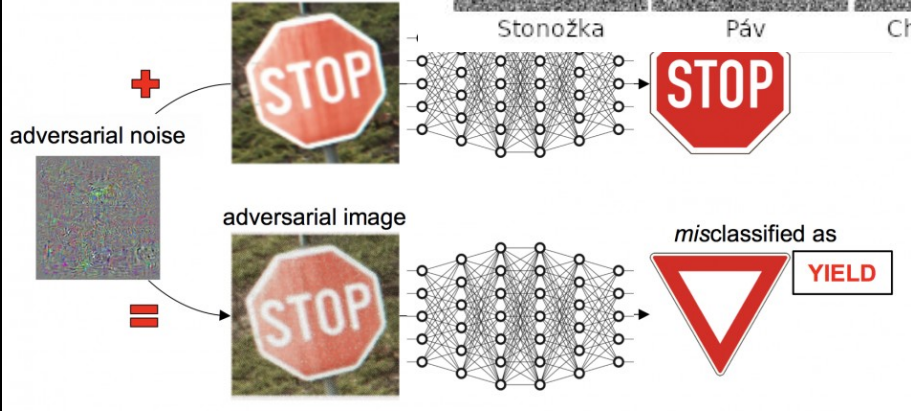
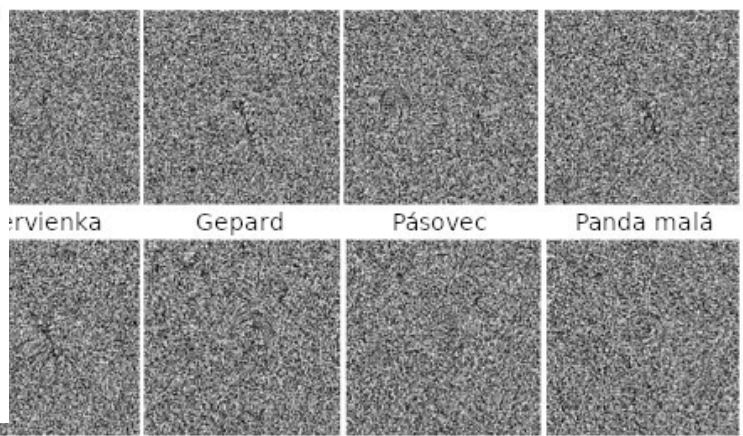
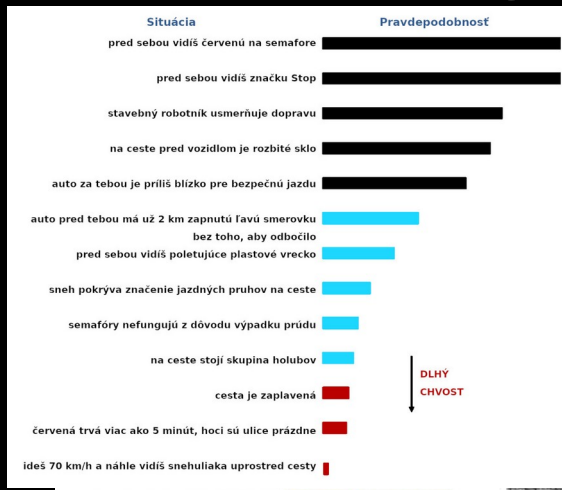
Prakticky nevieme, na základe čoho robia hlboké neurónové siete svoje rozhodnutia.

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“ hyperparametrov, „pohyby“ v modeloch, skrývanie dôvodov uvažovania, atď...

Pracujeme na vysvetliteľných a interpretovateľných systémoch AI ;-)

Technologické riziká a limity



1700

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**

Systémy AI skutočnú inteligenciu nemajú, len ju simulujú.

- **s vážnymi etickými dôsledkami**

Nie sú schopné rozlišovať morálne dobré a zlé!

Nechápu zmysel a dôsledky!



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

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!

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**

Neustály dohl'ad a jeho aplikácia...



CHINA'S SOCIAL CREDIT SYSTEM

It's been dubbed the most ambitious experiment in digital social control ever undertaken. The Chinese government plans to launch its Social Credit System nationally by 2020.

WHAT'S THE AIM?

The system intends to monitor, rate and regulate the financial, social, moral and, possibly, political behavior of China's citizens - and also the country's companies - via a system of punishments and rewards. The stated aim is to "provide the trustworthiness with benefits and discipline the untrustworthy."

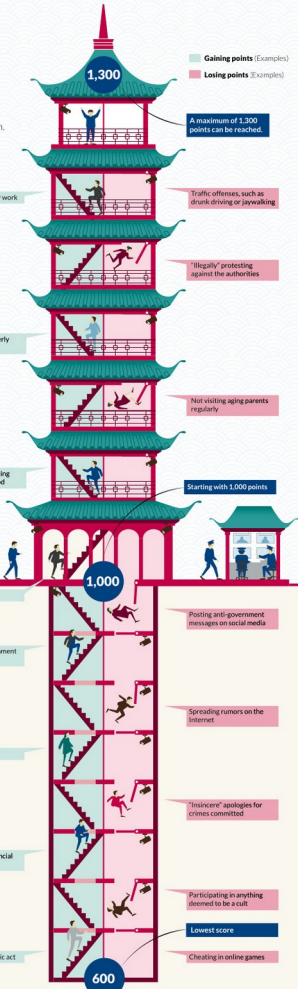
The Chinese government considers the system an important tool to steer China's economy and to govern society. There is still much speculation about how the final system will actually function. Details in this chart are based on pilot schemes and plausible expert expectations.

HOW DOES IT WORK?

Each citizen is expected to be given a social credit score that will increase or decrease depending on whether the subject's social behavior is acceptable.

The system is expected to draw on huge amounts of data about each and every individual, gathered from traditional sources - such as financial, criminal and government records and existing data from registry offices or school officials - along with digital sources. The latter include data collected on the internet, such as the subject's search history, shopping preferences on e-commerce sites and interactions on social media.

Moreover, the system could also rely on information obtained through video surveillance systems with help from facial recognition technology.



REWARDS AND PUNISHMENTS

Citizens with high scores get to enjoy special "privileges" while those with low scores ultimately risk getting treated as second-class citizens.

HIGH SCORES CAN LEAD TO

- Priority for school admissions and employment.
- Easier access to cash loans and consumer credit.
- Deposit-free bicycle and car hire.
- Free gym facilities.
- Cheaper public transport.
- Shorter wait times in hospitals.
- Fast-track promotion at work.
- Jumping the queue for public housing.
- Tax breaks.

PUNISHMENTS CAN LEAD TO

- Denial of licenses, permits and access to some social services.
- Exclusion from booking flights or high-speed train tickets.
- Less access to credit.
- Restricted access to public services.
- Ineligibility for government jobs.
- No access to private schools.
- Public shaming: exposure either online or on TV screens in public spaces of the names, photos and ID numbers of blacklisted citizens; phone dial tones mandated by authorities that inform people that they are calling a "disbowed debtor".

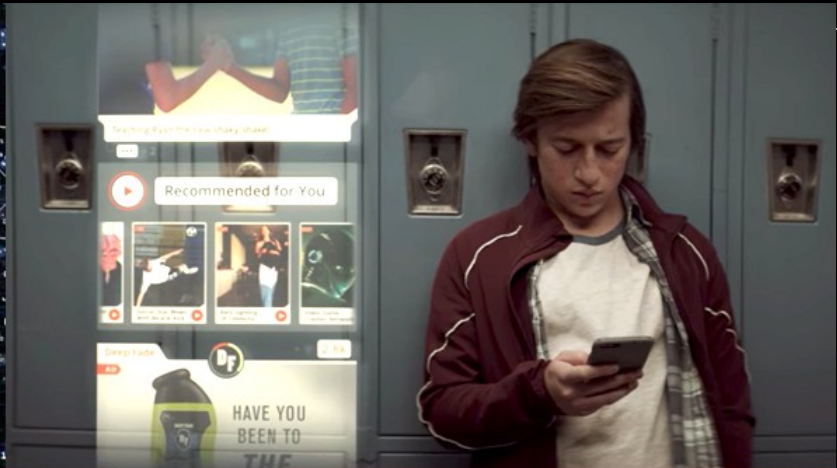
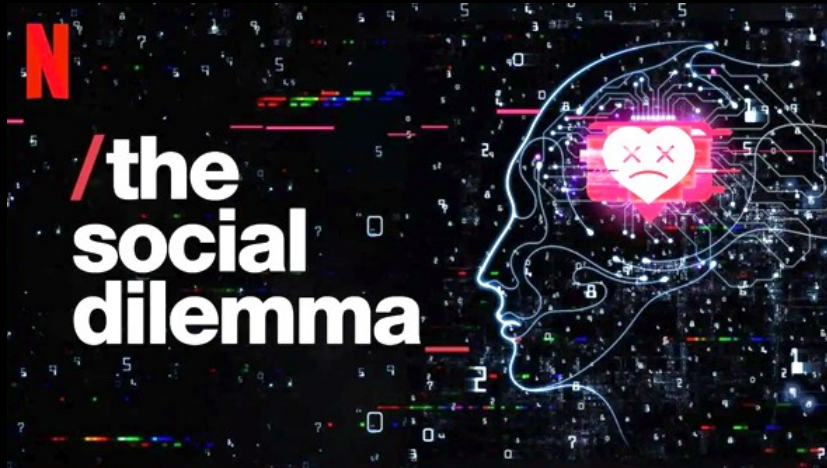
Prichádzame o vnútornú slobodu

- 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



Strácame kognitívne schopnosti

- riziko digitálnej demencie

Digitálna demencia



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...

Strácame kognitívne schopnosti

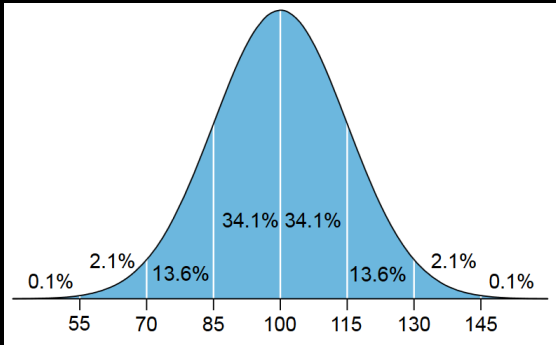
- riziko digitálnej demencie
- digitálne rozdelenie

Digitálne rozdelenie ako dôsledok AI



**INTELIGENČNÉ NOŽNICE,
KTORÉ SA OTVÁRAJÚ**

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



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...*

Strácame kognitívne schopnosti

- riziko digitálnej demencie
- 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

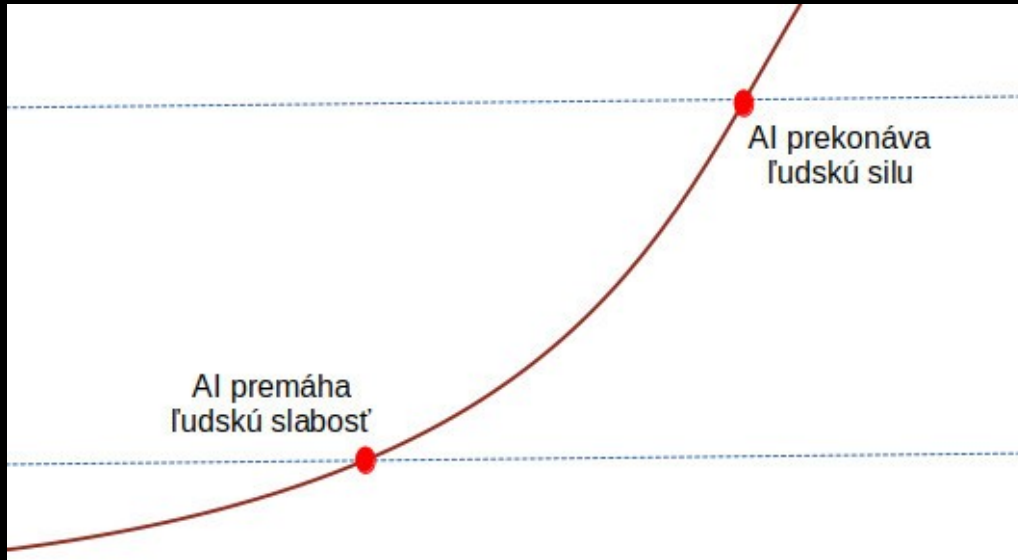


Dopad na **reálny** život a **skutočné** vzťahy

- **strata zmyslu** pre realitu a skutočné hodnoty
- **deepfake**
- virtuálne svety a **pokrivené** vzťahy
- **narušenie** psychického vývoja
- **deformácia** virtuálneho a informačného priestoru

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

Čoho sa reálne musíme obávať



Míľníkom, 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.

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

- **mat' 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 (dôvod: **veľká 4**)
 - adekvátne jednotlivým profesiám a spôsobu využívania systémov AI, **inak 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 vzdelávacej, vedecko-výskumnej, formačnej a administratívnej činnosti.

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**.

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*.**

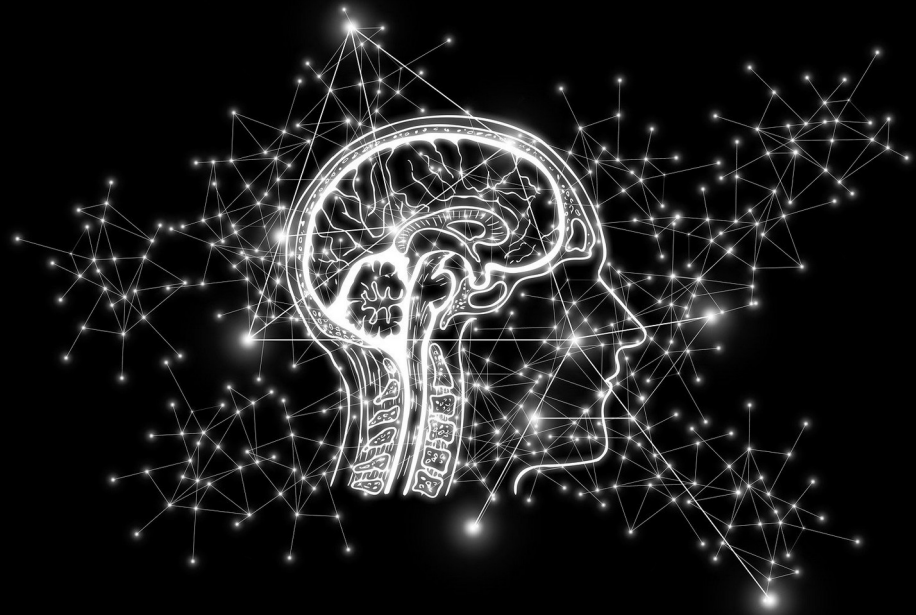
Je na nás, či umelá inteligencia bude dobrý sluha alebo zlý pán...

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; **dopomínová 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...



Ďakujem za pozornosť

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

