# Bezpečnost a etika v oblasti AI

Mgr. Igor Červený

Předmět: Využívání umělé inteligence (AI) jako nástroje pro práci VŠ pedagoga
Projekt: Podpora Rozvoje Učitelských Kompetencí (PRoUK)
Registrační číslo projektu: CZ.02.02.XX/00/23_019/0008385
Operační program: Jan Amos Komenský

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

# Úvod do etiky umělé inteligence

- Spojení etiky s technologiemi:
- Etika AI čerpá z poznatků etiky robotiky, automatizace a obecně etiky digitálních informačních a komunikačních technologií.

- Nové dimenze:
- Rychlý vývoj AI a její provázanost s dalšími technologiemi přináší nové etické výzvy, které nabývají na naléhavosti.

# Etické otázky a rizika AI – Bias a diskriminace

- Problematika biasu v datech: AI se učí z historických dat, což znamená, že pokud jsou data zkreslená nebo obsahují určité předsudky, AI tyto předsudky převezme a může je aplikovat při svých rozhodnutích.
- Například algoritmy pro nábor zaměstnanců mohou být zaujaté vůči určitému pohlaví nebo rase, pokud jsou taková data použita při tréninku modelu.

# Etické otázky a rizika AI – Bias a diskriminace

- Příklady diskriminace: Ve Spojených státech byly některé bankovní algoritmy obviněny z diskriminace černochů při schvalování úvěrů. Toto ukazuje, jak mohou být nedokonalá nebo zaujatá data příčinou nespravedlivých rozhodnutí.

# Etické otázky a rizika AI – Rostoucí spotřeba energie

- Generování odpovědí a multimédií pomocí umělé inteligence, zvláště u modelů hlubokého učení, vyžaduje obrovské množství energie. Například vygenerování jednoho obrázku je ekvivalentem nabití telefonu.
- Ekologická stopa: Za největšího znečišťovatele lze v tomto případě považovat infrastrukturu, datová centra a telekomunikační sítě. Pokud-li v brzké době nedojde ke změně obchodních a manažerských modelů, hrozí, že do roku 2040 dosáhne objem vyprodukovaných skleníkových plynů v oblasti ICT poloviny celkového objemu sektoru dopravy (BELKHIR, 2018).

# Etické otázky a rizika AI – Digitální (informační) a technologická propast

- Rozdíly v přístupu: Existují velké rozdíly v dostupnosti AI a souvisejících technologií mezi různými regiony, což prohlubuje digitální a technologickou propast.
- Ekonomické a sociální důsledky: Země s omezeným přístupem k technologiím jsou znevýhodněny, což vede k nerovnosti v oblasti vzdělávání, pracovních příležitostí a ekonomické prosperity.

# Etické otázky a rizika AI – Interkulturní aspekty AI (IBM)

- Kultury a zpracování dat: Algoritmy AI mohou obsahovat zkreslení a kulturní preference vyplývající z dat, na kterých byly trénovány. To může mít za následek nevyváženost nebo diskriminaci určitých kulturních skupin.
- Role firem jako IBM: Společnosti jako IBM se zaměřují na zohlednění interkulturních aspektů při vývoji a implementaci AI, aby se minimalizovala rizika spojená s kulturními rozdíly a zajistila se přístupnost pro různost kultur.

# Ochrana/Zábran a EU

- Evropské regulace: Evropská unie se zaměřuje na ochranu práv občanů při vývoji a používání AI prostřednictvím norem a předpisů, jako je GDPR nebo AI Act.
- Zákazy a omezení: Určité typy AI aplikací, však mohou být v rámci EU zakázány nebo přísně regulovány.

# Transparentnost a vysvětlitelnost AI

- Potřeba vysvětlitelnosti: V mnoha oblastech, jako je zdravotnictví nebo soudnictví, je nutné, aby byla rozhodnutí AI transparentní a vysvětlitelná. Pacienti potřebují vědět, proč AI doporučuje určitou léčbu, stejně jako soudce potřebuje pochopit, na základě jakých důvodů byl obviněný klasifikován určitým způsobem.
- Přístupy k vysvětlitelnosti: Techniky jako LIME (Local Interpretable Model-agnostic Explanations) a SHAP (SHapley Additive exPlanations) jsou používány k tomu, aby modely byly více vysvětlitelné a uživatelé mohli lépe rozumět rozhodnutím, která dělají.

# Transparentnost a vysvětlitelnost AI

- Černá skříňka: Mnoho moderních AI systémů, zejména těch založených na hlubokém učení, je považováno za „černé skříňky“, což znamená, že je obtížné pochopit, jak model dospěl ke konkrétnímu rozhodnutí. To je problematické v citlivých oblastech, kde je důležité rozumět důvodům rozhodnutí.
- Důsledky pro odpovědnost:
- Nedostatek transparentnosti ztěžuje přisuzování odpovědnosti.
- Potřeba zajistit, aby lidé mohli vysvětlit, proč AI rozhodla, jak rozhodla.

# Ochrana soukromí a údajů I.

- Slučování a využívání osobních údajů:
- AI pracuje s velkými daty a často vyžaduje shromažďování osobních údajů.
- Uživatelé si často neuvědomují, jak jsou jejich údaje shromažďovány, zpracovávány a sdíleny.

- Problematika transparentnosti:
- Nedostatečná informovanost o tom, jak se data využívají.
- Požadavky na etické shromažďování dat, respekt k soukromí a informovaný souhlas.

# Ochrana soukromí a údajů II.

- Etika evaluace:
- Shromažďování údajů pro evaluaci vyžaduje jasné souhlasy a transparentní komunikaci.
- Ve světě AI jsou této podmínky často obtížně dosažitelné.

- Příklady:
- Sociální sítě shromažďují data, jejichž účel je pro uživatele nejasný.
- Údaje získané v jednom kontextu mohou být využity v jiném bez vědomí uživatele.

# Ochrana soukromí a údajů III.

- Autorská práva:
- Generativní modely AI, trénované na rozsáhlých datových sadách, mohou vést k nejasnostem ohledně původu obsahu, autority a dodržování autorských práv, zejména při vytváření obrazových, textových či audiovizuálních materiálů.
- Použití obsahu vytvořeného generativními modely může zahrnovat rizika spojená s porušením práv duševního vlastnictví, což může organizacím způsobit finanční, právní a reputační škody, včetně nákladných soudních sporů.

# Ochrana soukromí a údajů IV.

- Erotika a pornografie :
- Rozvoj AI v oblasti generování realistických obrazů a videí, včetně neetického využití jako svlékání oblečení na fotografiích či tvorby intimních scén bez souhlasu, vyvolává závažné otázky ochrany soukromí a etických limitů.
- Využití deepfake technologií může vést k závažnému narušení důstojnosti a soukromí jednotlivců, včetně rizik spojených s vydíráním, šikanou či poškozováním pověsti.

# Erotika a pornografie

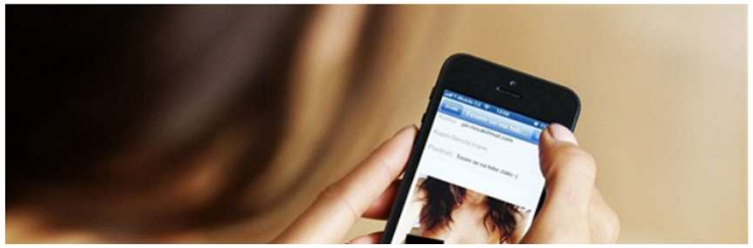




## Nahé fotky dívek šířili po sítích. Teenageři k tomu zneužili umělou inteligenci

21. září 2023 10:15

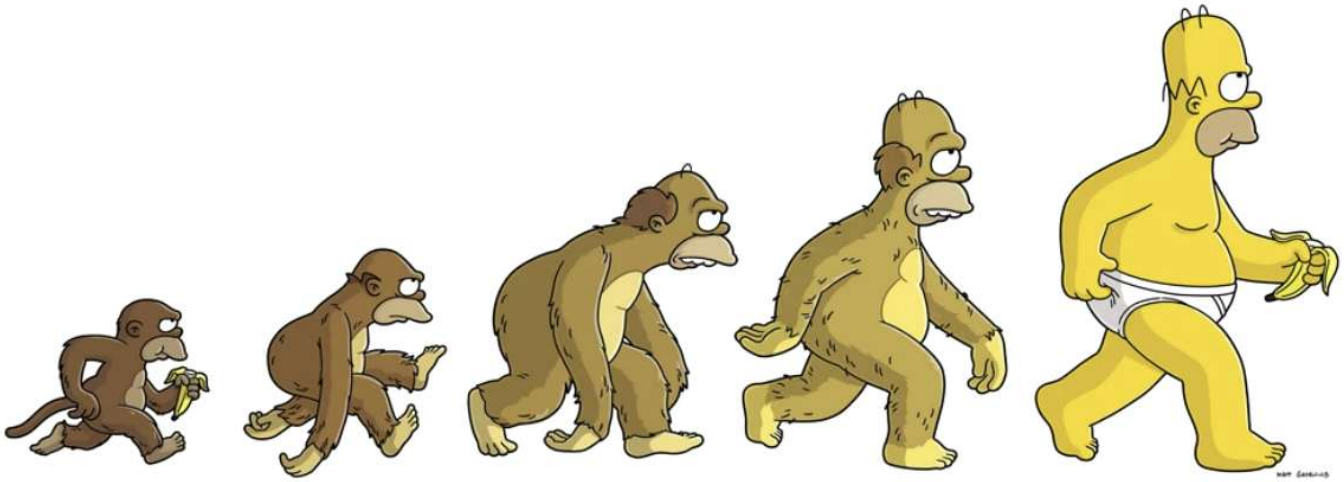
Španělská policie identifikovala 11 osob, které jsou podezřelé, že za pomoci umělé inteligence vytvořily falešné fotky nahých dívek z městečka Almendralejo na jihozápadě Španělska. Všichni podezřelí z vytvoření a šíření těchto fotografií na sociálních sítích jsou nezletilí, několika z nich je 12 a 13 let.




Zdroj: https://www.idnes.cz/zpravy/zahranicni/nahe-fotky-ai-spanelsti-teenageri-policie.A230921_100405_zahranicni_remy

# Lidé/AI a shromažďování dat

- *WEB 1.0*
- WEB 2.0
- WEB 3.0



# WEB 1.0

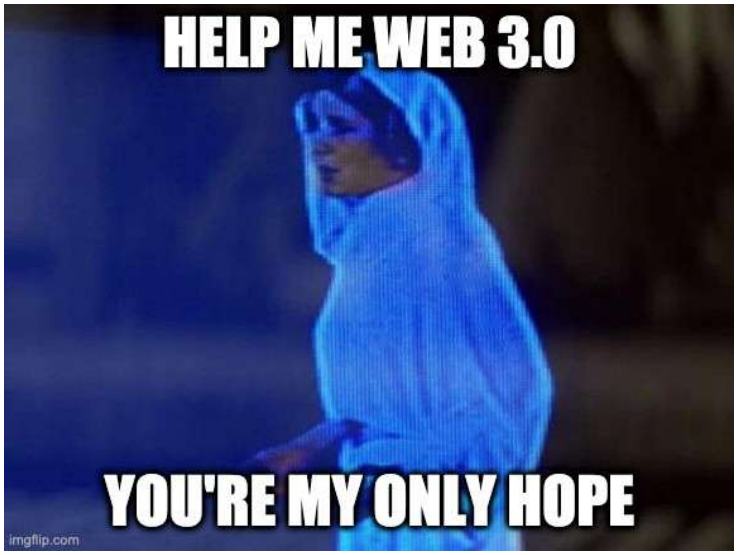
- Vznik Webu 1.0 se datuje od roku 1991, kdy Tim Berners-Lee vytvořil World Wide Web.
- Hlavními znaky Webu 1.0 jsou portály, statické HTML stránky s obsahem určeným výhradně ke čtení, obsahující možnost reagovat pouze skrze tzv. guestbooks (knihy návštěv).
- Počet uživatelů cca 45 milionů.


Zdroj:
https://www.earchiv.cz/b05/gifs/b0811004.gif

# WEB 2.0

- Vznik Webu 2.0 je řazen do roku 2004.
- Důraz je kladen na komunitu, kdy obsah vytvářejí sami uživatelé
  a navzájem jej sdílí skrze blogy a sociální sítě.
- Objevuje se používání jazyka XML, technologie AJAX, Flash, RSS a možnost tagování obsahu pomocí služeb jako je například Delicious.
- Počet uživatelů cca 5 miliard.

# Co je to WEB 3.0?



# WEB 3.0

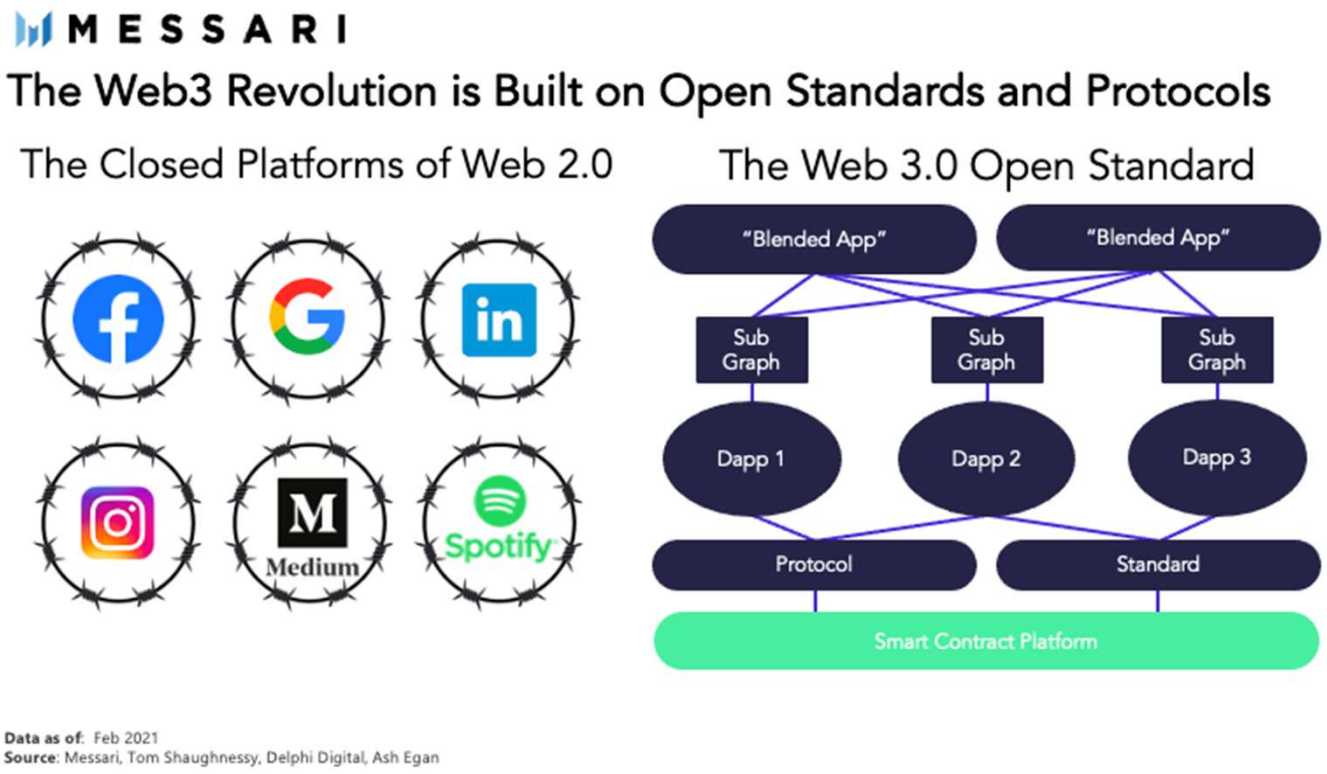
- Je zatím spíše teorie demokratičtějšího internetu, který funguje pro lidi a zároveň je vlastněný lidmi (nabourání paradigmatu, kdy „web“ vlastní jen několik vybraných organizací: FB, Google, YouTube apod).
- Ale až to přijde … bude to (velké):
  - stavět na interaktivních a sociálně propojených webech (decentralizace, VR a AR);
  - klást důraz na strojově usnadněné porozumění informacím na webu (AI, sémantický web);
  - poskytovat rámec pro opětovné použití a snazší sdílení informací mezi programy, komunitami a společnostmi (blockchain, krypto);

# WEB 3.0

| | web 1.0 | web 2.0 | web 3.0 |
|---|---|---|---|
| Vyhledávání | Katalog a hledání v katalogu. | Lokalizované vyhledávání pomocí klíčových slov nebo fráze. | Geografické vyhledávání položením dotazu ve formy věty s podmětem, předmětem, časem a místem. Personalizace podle uživatele. |
| Třídění a filtrace dat | Filtr volí uživatel. | Vyhodnocení geografické polohy a jazyka uživatele. Filtrace dat podle preference. | Personalizace podle povahy dat v kontextu k zobrazenému obsahu, místu, času a chování uživatele. |
| Reklama | Banner | Kontextuální k obsahu a poloze. | Personalizovaná s ohledem na obsah, místo a čas. |
| Forma prezentace | Webové stránky vs. aplikace na desktopu. | Webové aplikace s rozhraním na desktopu, přístup pro registrované účty. | Web OS, aplikace jako služba, přístup pod jednotnou identifikací. |
| Zdroj obsahu | Editor (administrátor). | Uživatel, podle povahy s/bez schválení editorem. | Schválená akce uživatelem během komunikace, např. záznam přednášky na mobil.<br>Událost generovaná zařízením, např. dopravní zácpa rozpoznaná GPS z průměrné rychlosti aut nebo ledničce dojde mléko. |

Zdroj: https://www.onlio.com/-/web-3.0

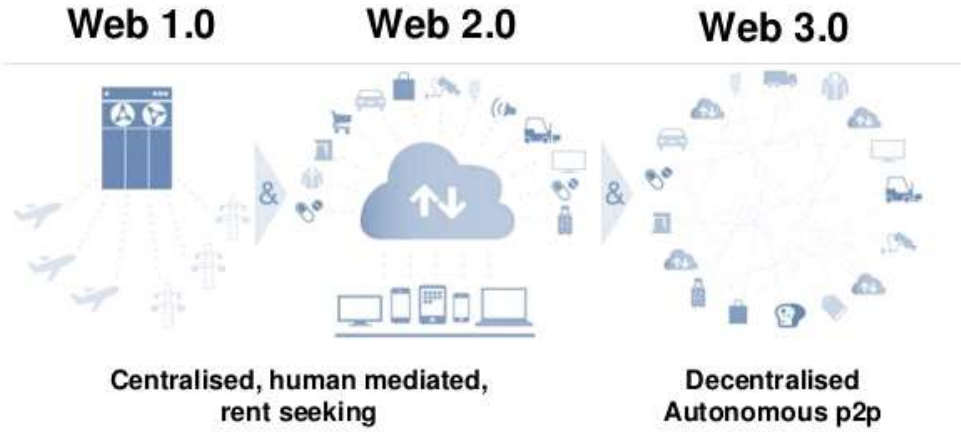
# Svět otevřených protokolů




Zdroj: https://messari.io/article/web3-eli5-what-is-web3

# Decentralizace

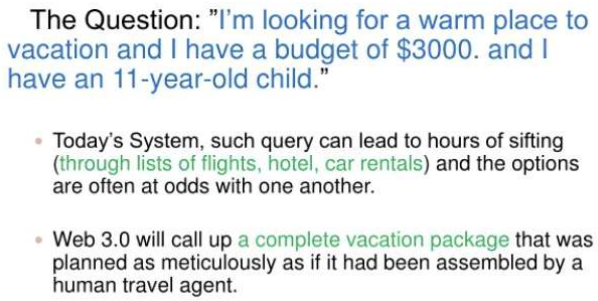

Zdroj:
https://www.reddit.com/r/Buttcoin/comments/rwca8k/connect_wallet/




Zdroj: https://medium.com/@pzoidis14/what-web-2-0s-journey-can-teach-us-about-the-eventual-adoption-of-web-3-0-3ab0bac0cd84

# Sémantický web

- Umožní vyhledávání pomocí významu vět, tedy rozpoznání podmětu, předmětu, jejich vzájemného vztahu se zřetelem na případné určení času a místa.



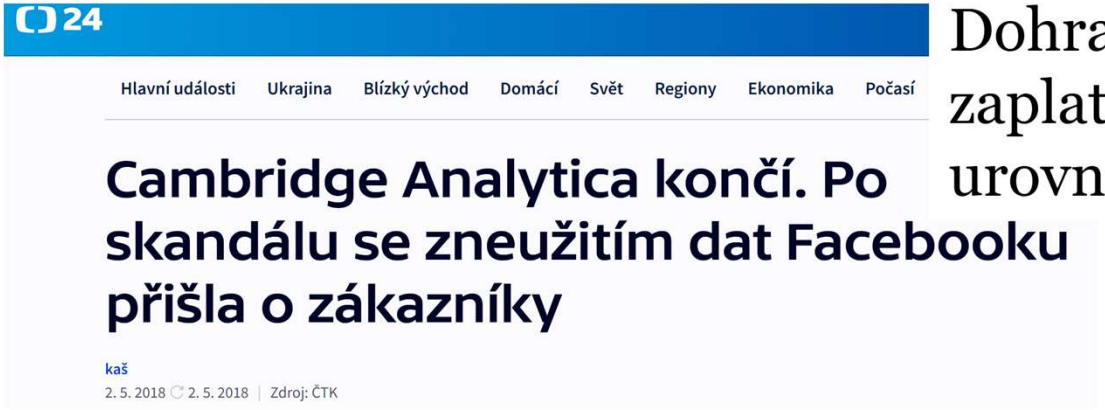

- Dále bude možné vyhledávačům „vysvětlit“, že slovo „špatný“ může v některých případech znamenat „dobrý“.

# Manipulace a zneužití AI I.

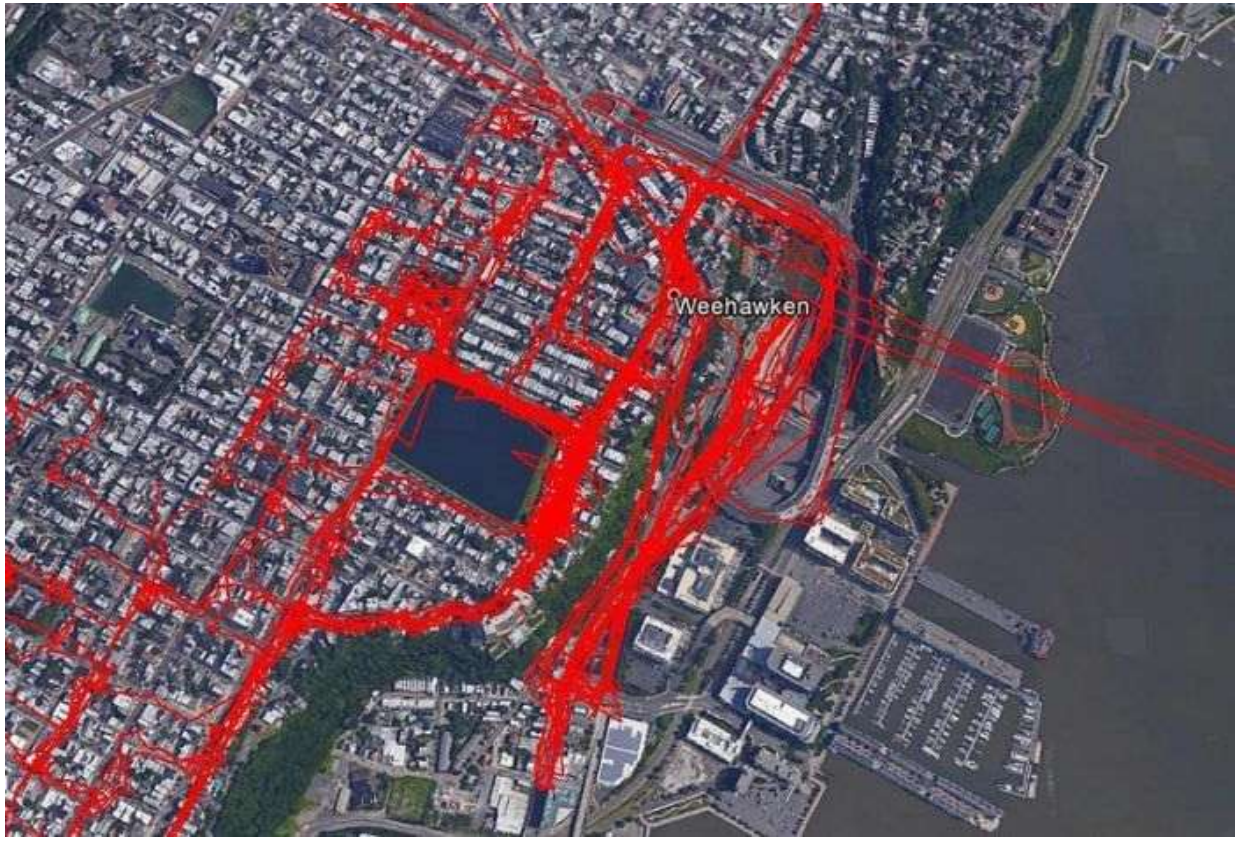
- Rizika manipulace:
- AI může ovlivňovat spotřebitelské chování i politické preference, protože „my lidí“ věříme obsahu zpráv, které sledujeme.
- Příklady: Cambridge Analytica, falešné zprávy (fake news) apod.

- Digitální vykořisťování:
- Uživatelé sociálních sítí produkují data, která jsou využívána bez jejich plného pochopení nebo finanční odměny.

# Cambridge Analytica

# Surveillance Capitalism




Zdroj: https://www.huffpost.com/entry/surveillance-capitalism-do-you-know-whos-tracking_b_59baa4f1e4b06b71800c3751

# Manipulace a zneužití AI II.

- Zranitelnost uživatelů:
- Ohroženější skupiny: děti, starší lidé, osoby s omezenými schopnostmi.

- Příklady: AI hračky, chatboty:
- *China's "educational" robots come with a level of intelligence equivalent to that of low-end smartphones.*


Zdroj:
http://www.szkekaidi.com/page116?article_id=22

# Fake news a dopad na společnost

- Postpravda:
- Technologie AI může zesilovat dezinformace a ohrožovat důvěru ve společnost.
- Příklady: Chatbot Tay, deepfake videa.

- Vliv na mezilidské vztahy:
- Technologie AI může oslabovat osobní vztahy a vytvářet iluzi společenství.
- Příklad: Zvýšená závislost na digitálních "společnících".

# Deepfake

- Páč lidé si to přáli: model, který mluví jako Miloš Jakeš
- https://www.umeligence.cz/blog/pac-lide-si-to-prali-model-ktery-mluvi-jako-milos-jakes

# Virtuální identita I.

- Specifikem virtuální identity je, že je určena uživatelem, tedy že uživatel rozhoduje o své (sebe)reprezentaci v kyberprostoru:
  - Od možnosti prezentovat se takový, jaký si myslí, že skutečně je;
  - Přes omezení se na prezentaci jen části sebe sama (např. jen lichotivé vlastnosti/postoje);
  - Anebo si přisoudit takové vlastnosti, o kterých ví, že je ve skutečnosti nemá (do virtuální identity se tak do značné míry mohou otiskovat naše vědomá i nevědomá přání, Šmahel, 2003a).

# Virtuální identita II.

- Specifikem virtuální identity je, že je určena uživatelem, tedy že uživatel rozhoduje o své (sebe)reprezentaci v kyberprostoru:
  - Umožňuje mu (stejně jako hra) experimentovat se samotnými prostředky a cíli, aniž by se musel (zcela a za určitých podmínek) podrobovat skutečným důsledkům „reálného“ chování, a přitom z toho mohl mít emocionální prospěch.

# Virtuální identita – jsem to ještě já?

- Prezentace Virtuální identity nám poskytla udržitelnou možnost pěstovat si permanentní kontrolu nad vlastní prezentací vůči okolí.
- To sebou nicméně přináší paradox:
  - „Autentické já“ reprezentované prostřednictvím digitálních médií se tedy současně stává „skutečnějším“ a „upravenějším“ než v prostředí tváří v tvář (Lüders, 2007).


zdroj: https://www.instagram.com/p/CZm6GibPyvS/

# Virtuální identita – důvěra/autenticita

## Youtuberka celou dobu lhala o svém vzhledu díky obličejovému filtru. Ve skutečnosti jí je 58 let

Díky omlazujícímu efektu vydělávala desítky tisíc korun.

zdroj: https://refresher.cz/67833-Youtuberka-celou-dobu-lhala-o-svem-vzhledu-diky-oblicejovemu-filtru-Ve-skutecnosti-ji-je-58-let

Zdroj: https://twitter.com/MattNavarra/status/1166746075857862656/photo/1

# Virtuální identita – důvěra/autenticita

- Možnosti a důsledky vyžívání internetu v oblasti experimentování s vlastní identitou (vědomé zkoušení být někým jiným*), či alespoň její „vylepšení“ (např. app pro automatickou úpravu fotek, úprava komentářů, smazání předešlých postů apod.),s sebou nicméně přinesly i řadu problémů v chápání autenticitě identity:
- když vím, co dělám já (mlžím/lžu), tuším, že to můžeš dělat i ty!

- *jen málo lidí se skutečně zapojilo do exotických experimentů (Parks a Roberts, 1998)

# Virtuální identity



Zdroj: https://bw.glbnews.com/01-2020/52780566523431/

# Etika a budoucnost psaní (nejen na školách)

- Generativní modely umělé inteligence zásadně ovlivňují oblasti, jako je marketing, žurnalistika a další odvětví závislá na produkci obsahu, neboť umožňují vytvářet texty téměř nerozeznatelné od lidských:
- Což přináší problém nejenom v oblasti duševního práva ale především osobnostní/profesního rozvoje.

# Etika a budoucnost psaní (nejen na školách)

- S rostoucím využíváním generativní umělé inteligence ve vzdělávacím procesu vyvstává potřeba zavádět celoplošná a jednotná pravidla, která by stanovila rámec pro její etické, bezpečné a pedagogicky přínosné využití. Tato pravidla by měla být srozumitelná, předvídatelná a závazná jak pro žáky, studenty a rodiče, tak pro vyučující či výzkumné pracovníky.
- Východiskem může být škálovatelný model využívání AI, jak jej navrhla například Univerzita Karlova (FF UISK), který rozlišuje pět úrovní využití AI – od zcela bez AI až po plnou integraci AI jako spolupracujícího nástroje. Každá úroveň přitom odpovídá specifickým formám tvůrčího procesu a míře autonomie uživatele. Důležitou roli zde rovněž hraje deklarace použití AI, která přispívá k transparentnosti a rozvíjí u žáků/studentů schopnost reflexe vlastního tvůrčího procesu.
- Ve školním prostředí se tudíž přístup k AI může pohybovat v několika základních variantách:

# Etika a budoucnost psaní (nejen na školách)

- Úplný zákaz využití AI: V některých vzdělávacích kontextech, zejména u zkoušek, testů nebo při samostatné práci, může být AI zakázána. Cílem je ověřit vlastní porozumění žáka bez vnější asistence a zachovat autenticitu výstupu. Tento přístup vyžaduje jasné komunikační nastavení a vysvětlení důvodů zákazu.
- Omezené využití AI pro přípravu nebo inspiraci: AI může být využívána jako nástroj pro brainstorming, hledání námětů či návrh osnovy projektu. Finální práce však musí být plně zpracována žákem. Tento model podporuje využití AI jako „mentálního skicáku“ bez přebírání jejího výstupu.

# Etika a budoucnost psaní (nejen na školách)

- AI jako jazykový nebo stylistický korektor: V tomto přístupu může AI pomoci s úpravou pravopisu, gramatiky nebo stylistiky žákova textu. Nedochází k tvorbě nového obsahu, ale ke zlepšení srozumitelnosti vlastního sdělení. Výsledkem je rozvoj sebereflexe a komunikačních dovedností.
- AI jako spolupracující nástroj: Žák může využít AI k vytvoření vybraných částí projektu (např. shrnutí, návrh anketního formuláře, návrh titulní stránky prezentace apod.), přičemž klíčový obsah zůstává jeho autorským dílem. Tento přístup vyžaduje vedení k odpovědnosti a schopnosti výstupy AI zhodnotit, upravit a případně odmítnout.

# Etika a budoucnost psaní (nejen na školách)

- Plná integrace AI do výuky: V některých případech, zejména v předmětech zaměřených na informatiku, mediální výchovu nebo výuku o umělé inteligenci samotné, může být AI běžnou a doporučenou součástí pracovního postupu. V takovém režimu se AI používá jako nástroj k objevování, tvorbě i evaluaci a je předmětem reflexe ze strany učitele i žáků.
- !! Etická pravidla by neměla být pouze represivní, ale měla by žáky/studenty vést k rozvoji digitálních kompetencí, informační gramotnosti a mediálního myšlení !!

# Důvěra v AI

- Jak zvýšit důvěru ve fungování AI? Role transparentnosti a prevence předpojatosti.
- Je AI objektivní? Nebo zrcadlí lidské chyby a hodnoty?

- Vysvětlení vs. zdůvodnění:
- Může AI poskytovat morálně relevantní zdůvodnění?
- Jak "reprezentovat" hodnoty v AI?

- Kompromis mezi výkonností a vysvětelitelností:
- Jsou lepší predikce AI dostatečně eticky opodstatněné, pokud nemůžeme vysvětlit, jak k nim dospěla?

# Důvěra v AI

**Adam Bárta** • 2.
Šetřím čas a zvyšuji zisk. Pomáhám (nejen) malým firmám z...
5 d. •

Navázat spojení

🪴 AI jako zahradník? No... řekněme, že naše zahrada se teď vrátila do roku 2020 😅

Každý rok ta samá scéna.

Chodím po zahradě, v ruce nůžky, připravený na jarní střih.
A jako tradičně – vůbec si nepamatuju, co se má kdy stříhat.
Takže klasika: Google, články, obrázky, Youtube... a pak mě to letos napadlo:

„Proč se nezeptat ChatGPT?"

Ten mi naprosto sebejistě poradil, jak a co ustřihnout.
Já poslechl.

A výsledkem je... zahrada, která vypadá jako čerstvě po založení.
Půlka rostlin zmizela, druhá se tváří překvapeně. A třetí zatím mlčí, ale podezírám je, že spřádají plán pomsty 🌿

Ale co už – přece musím věřit technologiím, že jo? 😄
Tak si povíme za rok, jestli AI kromě tabulek a procesů zvládá i keře a rybízy.

📌 Ať žije experimentální zahradničení 2.0!

👉 Už jste taky někdy použili AI tam, kde by to jiného nenapadlo? Dejte mi vědět, ať se necítím sám 😁

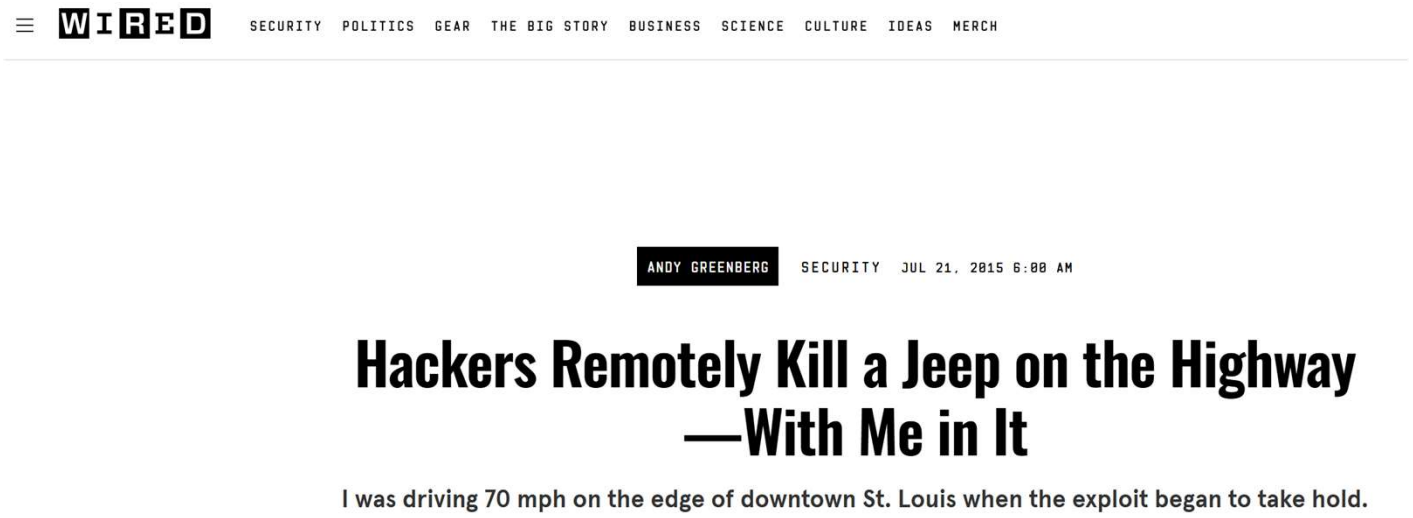
# Bezpečnost a zabezpečení

- Zranitelnost AI:
- Hacknutí autonomních vozů, kritické infrastruktury nebo zbraní může mít katastrofální důsledky.
- Rizika autonomních smrtících zbraní.

- Technologická zranitelnost:
- Více technologií = větší závislost a potenciální rizika selhání.
- Etická otázka kompromisů mezi efektivitou a bezpečností.

# AI a prolomení bezpečnosti

- Hackers Remotely Kill a Jeep on a Highway
- https://www.youtube.com/watch?v=MK0SrxBC1xs



# AI – Autonomní zbraně I.

Zdroj: https://www.seznamzpravy.cz/clanek/tech-techmix-autonomni-zbrane-potrebuji-regulaci-dohoda-je-zatim-v-nedohlednu-226914

Zdroj: https://www.denik.cz/veda-a-technika/armada-boj-robot-zabijeni-20221015.html

Zdroj: T2

# AI – Autonomní zbraně II.

- Born to be big:
- https://www.youtube.com/watch?v=DK6IGG5zRU8

# Informační bezpečnost I.

- Zabezpečení informací:
- zahrnuje fyzické, elektronické (virtuální) a duševní informace.
- platí po celou dobu životního cyklu informací, bez ohledu na formu jejich uložení.

- Tři pravidla informační bezpečnosti:
- 100% bezpečí neexistuje.
- Nejvíce problémů si způsobí každý sám.
- Prevence je účinnější než represe.

# Informační bezpečnost II.

- Cíle bezpečnostních mechanismů:

- Důvěrnost – Zajistit, že k datům mají přístup pouze oprávněné osoby.
- Dostupnost – Data jsou k dispozici, kdykoliv je to potřeba.
- Integrita – Zabezpečení přesnosti a ústnosti informací.
- Autentizace – Ověření identity osob a pravosti dat.
- Autorizace – Stanovení práv a přístupu v systému.
- Nepopiratelnost – Poskytnutí důkazu o provedení konkrétní operace.

# Technický a uživatelský pohled

- Technický pohled:
- Kryptologie, technické nástroje.

- Uživatelský pohled:
- Informační gramotnost, chování uživatelů.
- Datová a informační politika, informační audit.

- Digitální stopa:
- Vědomá stopa (e-maily, sociální sítě).
- Nevědmá stopa (cookies, fingerprinting).

# Kryptografie I.

- Nauka o šifrování a dešifrování informací.

- Cíle kryptografie:
- Zabezpečit citlivou komunikaci.
- Zajistit důvěryhodnost.
- Autorizace závažných úkolů (např. bankovní transakce).
- Ochrana před neoprávněným přístupem.

# Kryptografie II.

- Metody šifrování:

- **Transpoziční šifra**: Změna pořadí znaků.
- **Substituční šifra**: Nahrazení znaků (Caesarova šifra).
- **Symetrická kryptografie**: Stejný klíč pro šifrování i dešifrování.
- **Asymetrická kryptografie**: Veřejný a soukromý klíč.
- **Hašovací funkce**: Ochrana před poškozením dat.

# Základní procesy zabezpečení

- Identifikace: Určení identity subjektu.
- Autentizace: Ověření pravosti identity pomocí hesel, biometrie, čipů.
- Autorizace: Definice práv a oprávnění k přístupu.
- Nepopiratelnost: Vyloučení popření dřívější operace.

# Etické problémy a odpovědnost

- Kdo nese odpovědnost za rozhodnutí a činy, pokud jsou vykonány umělou inteligencí (AI)?
- Co znamená přisuzovat odpovědnost technologiím?

- Význam tématu:
- Roste nasazení AI v rozhodovacích procesech, které mají etické dopady.
- Komplexita AI systémů komplikuje jednoznačné přisouzení odpovědnosti.

# Aristotelský přístup k odpovědnosti

- Podmínky morální odpovědnosti podle Aristotela:
- Kontrola: Jednání musí mít původ v aktérovi.
- Epistemická podmínka: Aktér musí vědět, co dělá a jaké to může mít důsledky.

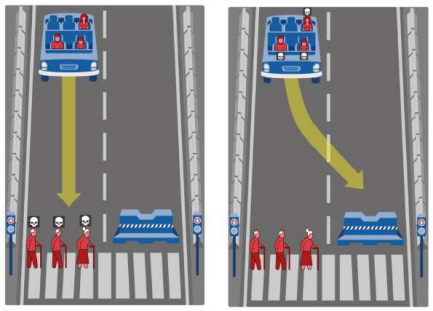
- Aplikace na AI:
- AI postrádá vědomí, svobodnou vůli a emoce, proto nemůže nést morální odpovědnost.
- Odpovědnost za činy AI by měli převzít lidé, kteří AI vyvíjejí nebo používají.

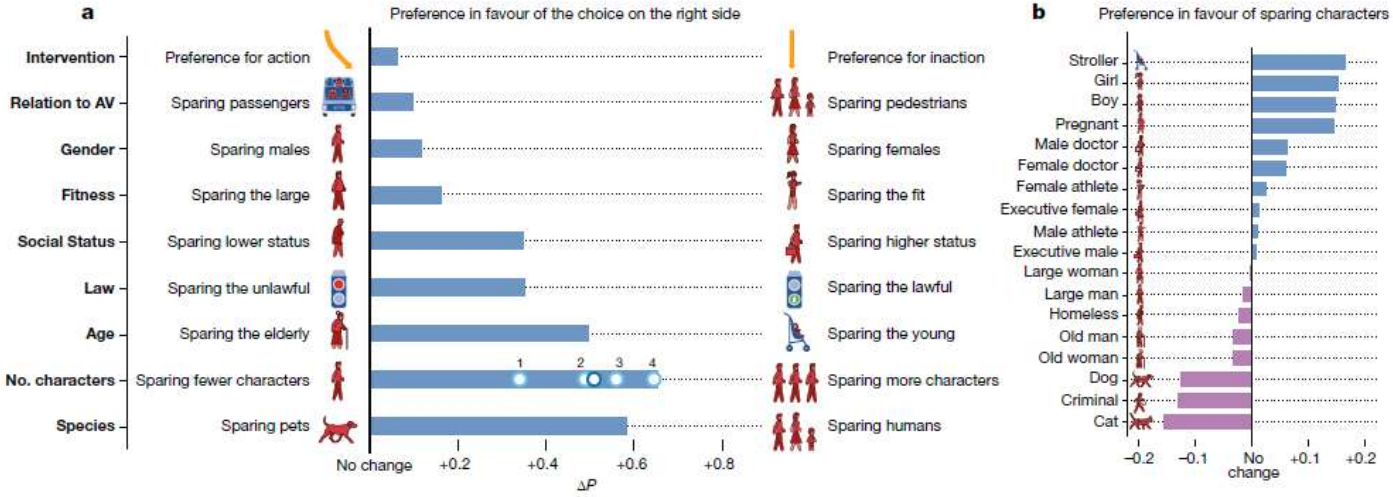
# Praktické problémy přisuzování odpovědnosti

- Rychlost rozhodování AI:
- U vysokofrekvenčního obchodování nebo autonomních vozidel nemá člověk čas zasáhnout.
- Historie AI systémů:
- AI může projít řadou aplikací (univerzita → zdravotnictví → vojenství). Kdo nese odpovědnost?
- Na výběru dat, tréninku a implementaci se podílí mnoho osob.
- Příklad: Autonomní automobil způsobí nehodu. Kdo je odpovědný: programátor, uživatel, automobilka, nebo regulační orgán?

# Přenecháme AI v oblasti rozhodování vše?

- The Moral Machine experiment:
- Platforma shromáždila 40 milionů rozhodnutí v 10 jazycích od milionů lidí ve 233 zemích a územích.





# AI a (autonomní) rozhodování

- Tesla Crashes Into Oncoming Car, Saves the Life of a Pedestrian!
- https://www.youtube.com/watch?v=Bh13LyX1q9M&ab_channel=ExploreAIHub

# Minimalizace újmy

- Definice: Minimalizace újmy se zaměřuje na snižování negativních dopadů AI na společnost, jednotlivce a přírodu.

- Etický rámec:
  - Zásada prevence: Předcházet předvídatelným škodám.
  - Princip proporcionality: Vyvažování přínosů a rizik.
- Praktické aspekty:
  - Odstraňování předsudků v datech a modelech.
  - Zajištění robustnosti systémů.
  - Transparentnost algoritmů.

# 5 základních etických principů AI I.

- Princip neškození:
- Tento princip vychází z povinnosti zajistit, aby systémy umělé inteligence nepoškozovaly jednotlivce, skupiny ani společnost. To zahrnuje prevenci fyzického, psychického či sociálního poškození, včetně diskriminace, ztráty soukromí či šíření dezinformací.

- Princip prospívání:
- Cílem tohoto principu je zajistit, aby AI přinášely užitek pro jednotlivce i společnost. To zahrnuje podporu inovací, zlepšování kvality života, ochranu životního prostředí a spravedlivé rozdělování výhod, které tyto technologie nabízejí.

# 5 základních etických principů AI II.

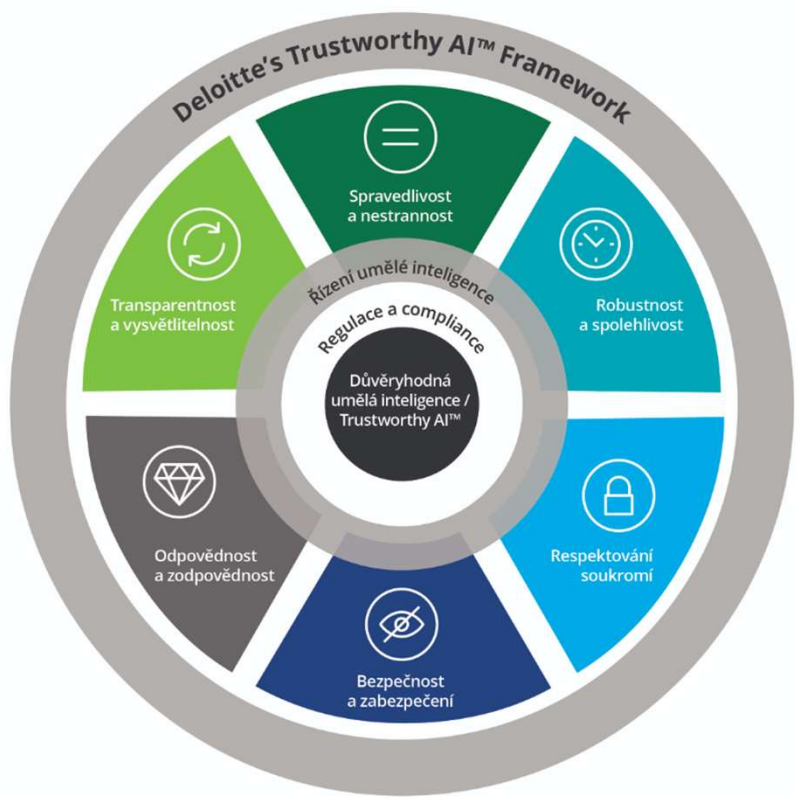
- Princip respektu k lidské autonomii:
- Respektování lidské autonomie znamená, že AI nesmí narušovat schopnost lidí činit vlastní rozhodnutí. Uživatelé musí mít kontrolu nad tím, jak se systémy AI používají, a jejich volby musí být informované, svobodné a respektované.

- Princip spravedlnosti:
- Spravedlnost zahrnuje rovné zacházení se všemi uživateli a zamezení diskriminace na základě pohlaví, věku, rasy, náboženství nebo jiných charakteristik. AI by měla být navržena a nasazena tak, aby její výhody i zátěže byly rozděleny rovnoměrně.

# 5 základních etických principů AI III.

- Princip vysvětlitelnosti:
- Vysvětlitelnost vyžaduje, aby byly algoritmy a rozhodnutí systémů AI srozumitelné pro (téměř každého) uživatele. Tento princip je klíčový pro budování důvěry, umožňuje ověřitelnost rozhodovacích procesů a přispívá k odpovědnosti za výsledky.

# Důvěryhodná umělá inteligence a její potenciál – dle Deloitte




Zdroj:
https://www2.deloitte.com/cz/cs/pages/risk/solutions/trustworthy-ai.html

# 1. Transparentnost a vysvětlitelnost

- Důvěra v AI vyžaduje schopnost porozumět jejím nedostatkům, identifikovat zkreslení, provádět kontroly přesnosti systémů a po nasazení modelů sledovat jejich změny. Toho lze dosáhnout například prostřednictvím modelování výstupů, trasováním nebo průběžným monitorováním. Každý z těchto přístupů má vliv na účinnost systému.

# 2. Spravedlnost a nestrannost

- Nestrannost spočívá v mnoha ohledech na kvalitě údajů, na kterých je model AI vyškolen. Podniky jsou pověřeny zajistit, aby datové sady neobsahovaly skryté zkreslení, aby se po nasazení posuzovala data z domén v reálném světě a aby součástí procesů byly interní a externí kontroly, které nepřetržitě monitorují a ověřují spravedlivost a ověřování AI.

# 3. Robustnost a spolehlivost

- Konzistence systémů je klíčovým faktorem úspěchu. Důvěryhodné systémy AI se musí chovat spolehlivě a podle očekávání, a to i v případě, že se setkají s neočekávanými daty. Měly by být také dostatečně robustní, aby zůstaly spolehlivé ve velkém měřítku. A to i z bezpečnostního hlediska – systém by měl být odolný takovým způsobem, aby dokázal čelit potenciálním snahám oklamat model, kvůli kterým by následně dodával nepřesné výstupy.

# 4. Bezpečnost a zabezpečení

- Organizace by měly nejen stavět na systémech, které jsou odolné vůči známým (a neznámým) rizikům, ale také by měly informace o těchto rizicích sdělovat uživatelům systémů a dalším zúčastněným stranám. V praxi to může znamenat školení modelů, využívání vrstvených procesů s redundancí zabudovanou do infrastruktury a přísné řízení a podávání zpráv, které prostupuje celým životním cyklem AI.

# 5. Odpovědnost a zodpovědnost

- Za každým nástrojem AI stojí řada lidí, kteří rozhodují o tom, jak je systém konstruován a používán v praxi. Důvěryhodná umělá inteligence je potom založena na lidské zodpovědnosti za chování takového nástroje. Jedná se o jednu z největších výzev v oblasti AI a její řešení vyžaduje jasný řetězec odpovědnosti za nápravu problémů, které se projevují v provozu.

# 6. Respektování soukromí

- U důvěryhodné AI by organizace měly zajistit, aby byla chráněna trénovací data, vstupy byly omezeny jen pro vlastníka modelu, výstupy byly dodávány pouze uživateli a samotný model byl chráněn před nežádoucí expozicí. V praxi uživatelé potřebují jasné informace, aby pochopili, jaká data jsou používána a jakým způsobem. Dále musí mít uživatelé možnost volby se ke sdílení dat přihlásit a také se od něho odhlásit. A v neposlední řadě je potřeba dostupný kanál pro sdělování zpětné vazby.

# Otázky / Debata

# Děkuji za Vaši pozornost

Mgr. Igor Červený

Spolufinancováno Evropskou unií

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY