Máme se BÁT umělé inteligence?

The Karel Čapek Center
for Values in Science and Technology

Umělá inteligence je všudypřítomná

Umělá inteligence (artificial intelligence, AI), přesněji umělé systémy (počítačové programy) vykazující inteligentní chování, se zdá být všudypřítomná. Rozhoduje o poskytnutí půjček, povýšení či propuštění, v reálném čase rozpoznává lidské tváře. Diagnostikuje choroby, čte rentgenové snímky, hledá nová léčiva a usnadňuje komunikaci mezi zdravotníky a pacienty. Řídí auta, letadla a další dopravní prostředky, pomáhá distribuovat energii. Předkládá obsahy na sociálních sítích, pohání vyhledávací nástroje v prohlížečích jako je Bing či Chrome, sleduje naše jednání online a nabízí personalizovanou reklamu. Je „mozkem“ stále rostoucího počtu typů robotů, včetně záchranářských či vojenských. Hledá nejlepší trasu, řídí složité logistické operace, sumarizuje texty a generuje vlastní, včetně básní a povídek, skládá hudbu, produkuje velmi povedené obrázky a umělecká díla. Vede s námi konverzaci v přirozeném jazyce, překládá i velmi složité texty, kontroluje gramatiku a stylistiku, připravuje za nás odpovědi na maily, automatizuje celou řadu procesů.

Umělá inteligence nepochybně představuje mimořádně užitečný nástroj, který může zefektivnit naši práci a poskytnout nám nové možnosti aktivního i pasivního odpočinku. Může nás vzdělávat, informovat, bavit a celkově zlepšovat kvalitu lidského života. Vývoj a využívání umělé inteligence, alespoň v našem západním kulturním prostředí, byly však vždy spojeny s obavami a katastrofickými scénáři. Mnohým z nás se může vybavit záludný počítač HAL 9000 z kultovního snímku Stanleyho Kubricka 2001: Vesmírná odysea, který se pokusí vyhubit posádku kosmické lodi Discovery, nebo slavný Terminátor, poslaný zpět v čase umělou inteligencí známou jako Skynet, aby zabil Sarah Connorovou a zajistil si tak vítězství nad lidstvem, které ještě vzdoruje snaze o naprostou eliminaci.

Umělá inteligence, ať již v podobě počítačových systémů nebo robotů, má zkrátka špatnou pověst. A v posledních letech se objevuje celá řada technopesimistů, kteří se domnívají, že tato pověst může být zcela oprávněná. Umělá inteligence podle nich představuje obrovskou, možná dokonce největší hrozbu pro lidskou společnost. V roce 2022 se začaly objevovat nové a v jistém smyslu revoluční systémy umělé inteligence, které překvapují a fascinují svými schopnostmi. Program GPT-4 dokáže konverzovat v přirozeném jazyce a pozoruhodně dobře plnit celou řadu lidských příkazů, jiné programy, jako je Midjourney, umějí generovat úžasné obrázky na základě slovních popisů. V souvislosti s tímto nesporným pokrokem se s novou intenzitou vynořují staré obavy o umělé inteligenci jako posledním lidským výtvoru, protože až převezme vládu nad světem, pro lidi v něm již nemusí být místo.

Myslíme si, že tyto obavy jsou přehnané. Umělá inteligence, alespoň v současné podobě, není sama o sobě žádným zásadním rizikem, jak se zde pokusíme ukázat. Mnohem větším a především aktuálním, nikoli pouze spekulativním rizikem jsou způsoby, jak umělou inteligenci využíváme a jak ji trénujeme. Kromě toho jí někdy dostatečně nerozumíme a zatím jsme si nedokázali vypěstovat důležité návyky a postoje (shrnujeme je pod hlavičkou algoritmické gramotnosti), které by nám umožnili využívat moderních nástrojů a systémů umělé inteligence efektivně, bezpečně a ve prospěch nás všech. Jinými slovy, problémem a rizikem není umělá inteligence jako taková, ale naše pasivita a neschopnost transformovat vzdělávací a aplikační systém takovým způsobem, aby nám umožnily držet krok s moderním technologickým vývojem.

Co vlastně je umělá inteligence?

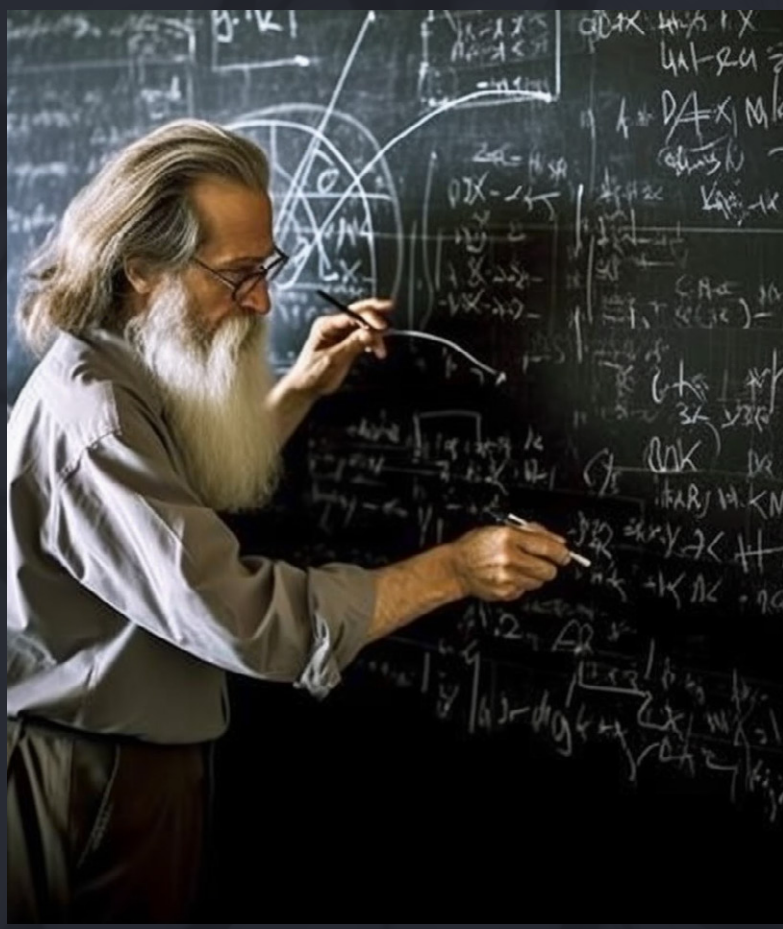
Vědecká disciplína a vlastnost umělých systémů

Výraz „umělá inteligence“ poprvé použil v roce 1955 americký informatik a kognitivní vědec John McCarthy, jeden z prvních průkopníků vědecké disciplíny, která se pokouší vytvořit inteligentní chování v umělých systémech. Tato relativně nová disciplína přijala název umělá inteligence a představuje mezioborovou vědu zahrnující přístupy moderní matematiky, kognitivní vědy, neurobiologie, informatiky, filosofie a dalších. Umělou inteligenci ale také chápeme jako vlastnost lidských výtvorů, které dokáží jednat způsoby, které považujeme za inteligentní. Podobně jako je lidská inteligence vlastností lidí, je umělá inteligence vlastností některých systémů, které se od sebe liší formou inteligence, její sofistikovaností i účely, pro které je vyvíjená a využívaná v praxi. Ptáme-li se, zda je (či může být) umělá inteligence rizikem, máme na mysli to, zda jsou či mohou být nějakým rizikem umělé systémy vykazující inteligentní chování.

rámeček č. 1

ALGORITMUS A PROGRAM

Algoritmus je abstraktní procedura, která poskytuje jednoznačný návod, jak pro nějaké vstupy dojít k odpovídajícím výstupům. Algoritmem je tak například návod, jak násobit dvě čísla. Programy jsou v nějakém jazyce (např. C++, Python, Java apod.) zapsané sledy instrukcí, které po spuštění programu provádí počítač a pro dané vstupy vytvoří („vypočítá“) příslušné výstupy. Programy můžeme chápat jako „ztělesnění“ algoritmů stroji.

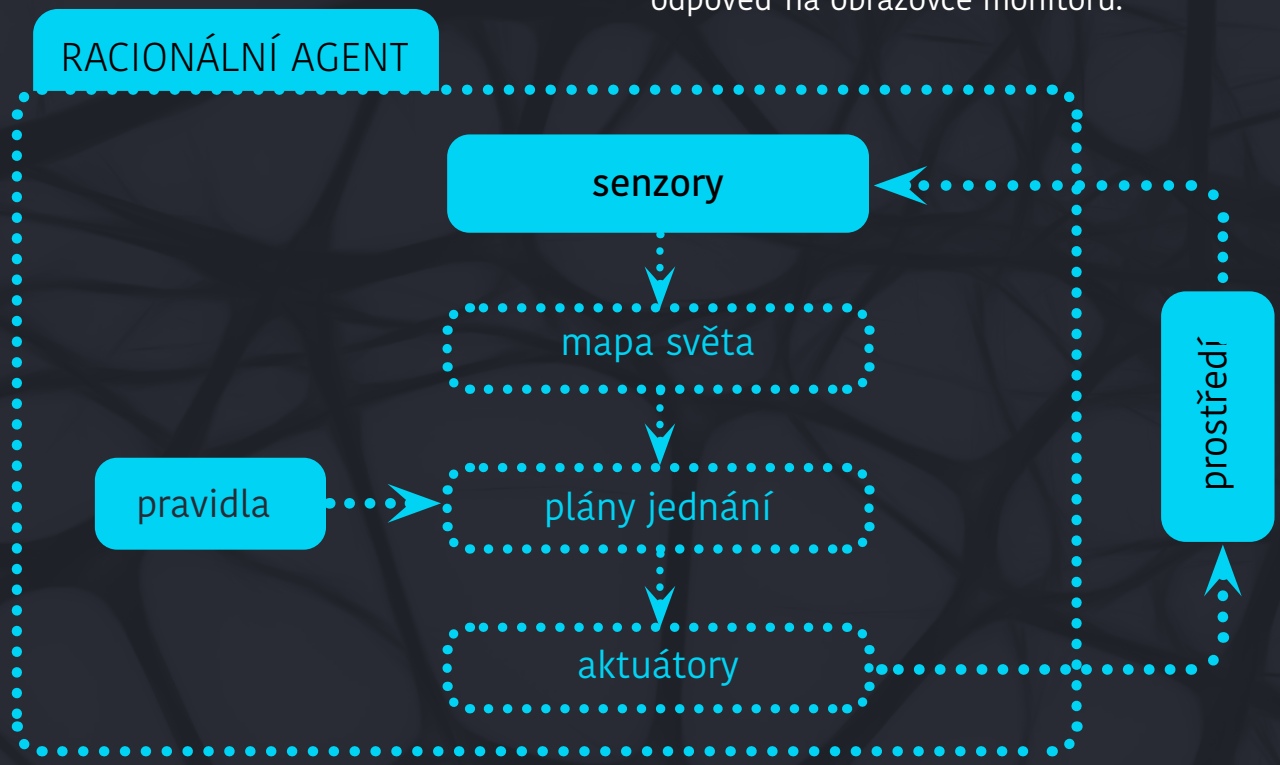


John McCarthy

Umělá inteligence jako vědecká disciplína je poměrně mladá a samotné chápání významu slova „inteligence“ se postupně měnilo. Nejdříve se spojovalo s jednáním, které se blíží lidskému a v konečném důsledku je od něj neodlišitelné (např. prostřednictvím tzv. Turingova testu), později se začal klást důraz na logickou strukturu myšlení (inteligentní myšlení postupuje ve shodě s principy logických systémů, v nichž je možné ho zapsat a vtělit do strojů). V poslední době se výzkum na poli umělé inteligence zaměřuje na vývoj tzv. racionálních agentů, kde se inteligentní chování chápe jako takové chování, které „rozumně“ řeší nějaké úlohy (a rozumně již neznamená „stejně či podobně jako člověk“). Jinými slovy, racionální agent je systém, který jedná takovým způsobem, aby dosáhl nejlepšího výsledku (či v případě nejistoty nejlepšího očekávaného výsledku).

Na moderní systémy umělé inteligence můžeme pohlížet jako na racionální agenty a na snahu „vtělit“ jim inteligenci jako na pokusy „přimět“ je jednat racionálně. Základní struktura takového agenta je velmi jednoduchá:

V principu takový systém dokáže prostřednictvím senzorů vnímat své prostředí, vytvářet si obraz světa kolem sebe, zpracovávat tyto poznatky a vytvářet plány jednání, a konečně prostřednictvím výkonných prvků (aktuátorů) jednat. Agenti mohou být reální, v reálném světě (roboti), ale i čistě virtuální (software). I velké jazykové modely, jako je GPT-4, lze chápat jako agenty: vnímají (virtuální) prostředí (naše dotazy), dokáží je zpracovat, ve shodě s tím, co se naučily vytvoří plán jednání (odpověď) a nakonec reagují se světem – zobrazí odpověď na obrazovce monitoru.



Neuronové sítě a hluboké učení

Při klasickém programování využívají programátoři algoritmy, které zapisují v nějakém jazyce, jemuž počítač rozumí. Počítač si posléze příkazy tohoto jazyka převádí na instrukce pro procesor. Magie moderních počítačů se zakládá na schopnosti procesoru načítat velká množství dat z paměti, neuvěřitelně rychle s nimi provádět logické operace a zase je ukládat do paměti:

KLASICKÉ PROGRAMOVÁNÍ

Nutnou podmínkou pro smysluplnou práci standardních počítačů je to, že každému počítači musíme zadat program, vytvořený na základě přesné znalosti postupu, jenž vede k řešení dané úlohy. Bez znalosti tohoto postupu nelze program vytvořit („naprogramovat“). Příkladem mohou být např. dlouhodobé, ale neúspěšné snahy o vytvoření algoritmů pro překladače mezi živými jazyky, snaha o počítačové autonomní systémy rozhodování na základě skenování okolního prostředí a pokusy o řešení mnohých dalších projevů inteligentního chování. Tyto neúspěšné snahy realizovat inteligentní chování v rámci standardního algoritmického postupu pro současné počítače vedly k vývoji velkého množství alternativních metod. Zvláštní pozornosti se v posledních letech dostává metodám strojového učení. Jejich podstatou je to, že algoritmy nejsou vstupem celého procesu (ve formě programů), ale jeho výstupem. Umělý systém tak na základě dat a učení vytváří algoritmy, které posléze řeší úkoly, pro něž jsou určeny:

STROJOVÉ UČENÍ

Algoritmy jsou samozřejmě nezbytné i během procesu učení (podle nich se systém učí), nejsou však používané k vytváření

výstupů (např. označení fotografie slovem „pes“), ale k formování pravidel, podle nichž posléze tyto výstupy vznikají.

rámeček č. 2

METODY UČENÍ NEURONOVÝCH SÍTÍ

V současné době se využívá několik základních metod, jimiž se neuronové sítě učí na souborech dat.

1. UČENÍ S UČITELEM (supervised learning)

K dispozici jsou sady označených dat a učitel (člověk) učí síť, jak k daným vstupům přiřadit správné označení. Daty mohou být například fotografie psů a koček (výstupem je tedy označení „pes“ a „kočka“). Prostřednictvím těchto dat a s pomocí lidského dozoru se síť naučí k fotografiím přiřadit správné označení. Jakmile proces učení skončí (síť je „natrénovaná“), je možné ji použít k vyhodnocování neoznačených fotografií.

2. UČENÍ BEZ UČITELE (unsupervised learning)

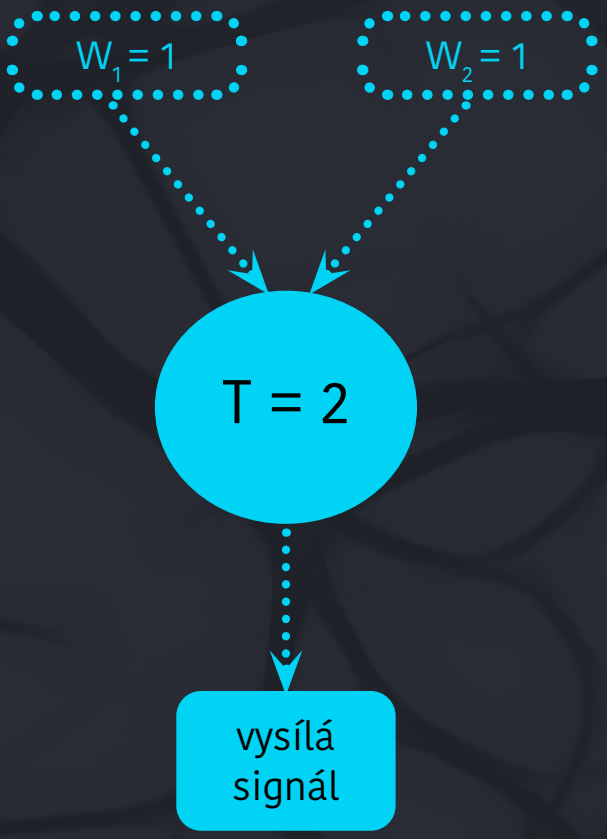
V tomto případě nejsou data nijak označená (je jich například moc, a proto je to v praxi těžko uskutečnitelné). Síť se učí rozpoznat v těchto datech nějaké vzory či zákonitosti.

3. POSILOVACÍ UČENÍ (reinforcement learning)

Síť má k dispozici data a učí se na základě odměny. Může se tak třeba naučit hrát počítačové hry. Například když náhodně zahraje „správný“ tah, získá odměnu (zpětnou vazbu) a postupně se tak naučí hrát hru velmi efektivně.

Důležitou roli v moderní umělé inteligenci založené na strojovém učení hrají umělé neuronové sítě. Umělé neuronové sítě se skládají z neuronů. Nejedná se však o reálné neurony, ale o konstrukce v programovacím jazyce, které jsou spojeny s nějakými parametry a funkcemi. Představme si například, že bychom chtěli naprogramovat neuron, který zvládá logickou operaci „a“, tj. operaci, která (zjednodušeně) pro dvě hodnoty „pravda“ poskytne hodnotu „pravda“ (a ve všech ostatních případech „nepravda“). Takový neuron musí přijímat podněty odjinud (musí „zjistit“, jaké pravdivostní hodnoty k němu doputovaly) a nějak se podle nich „rozhodnout“. Řekněme, že mu přiřadíme tzv. váhu, která představuje parametr, který spolurozhoduje o tom, jak bude neuron reagovat. V tomto případě je váha rovná jedné (pravda) nebo nule (nepravda). Dále mu musíme přiřadit nějaký práh; pod tímto prahem neudělá nic, pokud je dosažený či překročený, vyšle signál dál. V případě spojky „a“ je tímto prahem hodnota „2“. Když ke vstupům neuronu doputují hodnoty 1 a 1, sečte je a protože dosáhl prahu, vyšle signál. Ten nám říká, že pro dané vstupy je výstupem hodnota „pravda“. Kdyby ale třeba dostal vstupy 1 a 0, výsledná hodnota 1 by byla pod prahem a neuron by neudělal nic; věděli bychom, že v tomto případě je výsledná hodnota „nepravda“.

Tento neuron je poměrně jednoduchý a v moderních neuronových sítích se nacházejí neurony podstatně komplikovanější, jejichž parametry a interakce jsou dány sofistikovanými matematickými funkcemi.



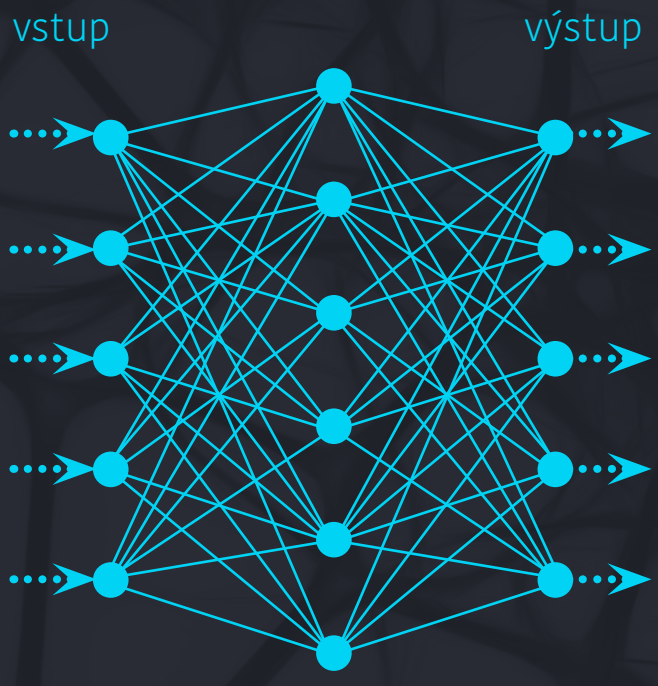
T = práh (treshold)

W = váha (weight)

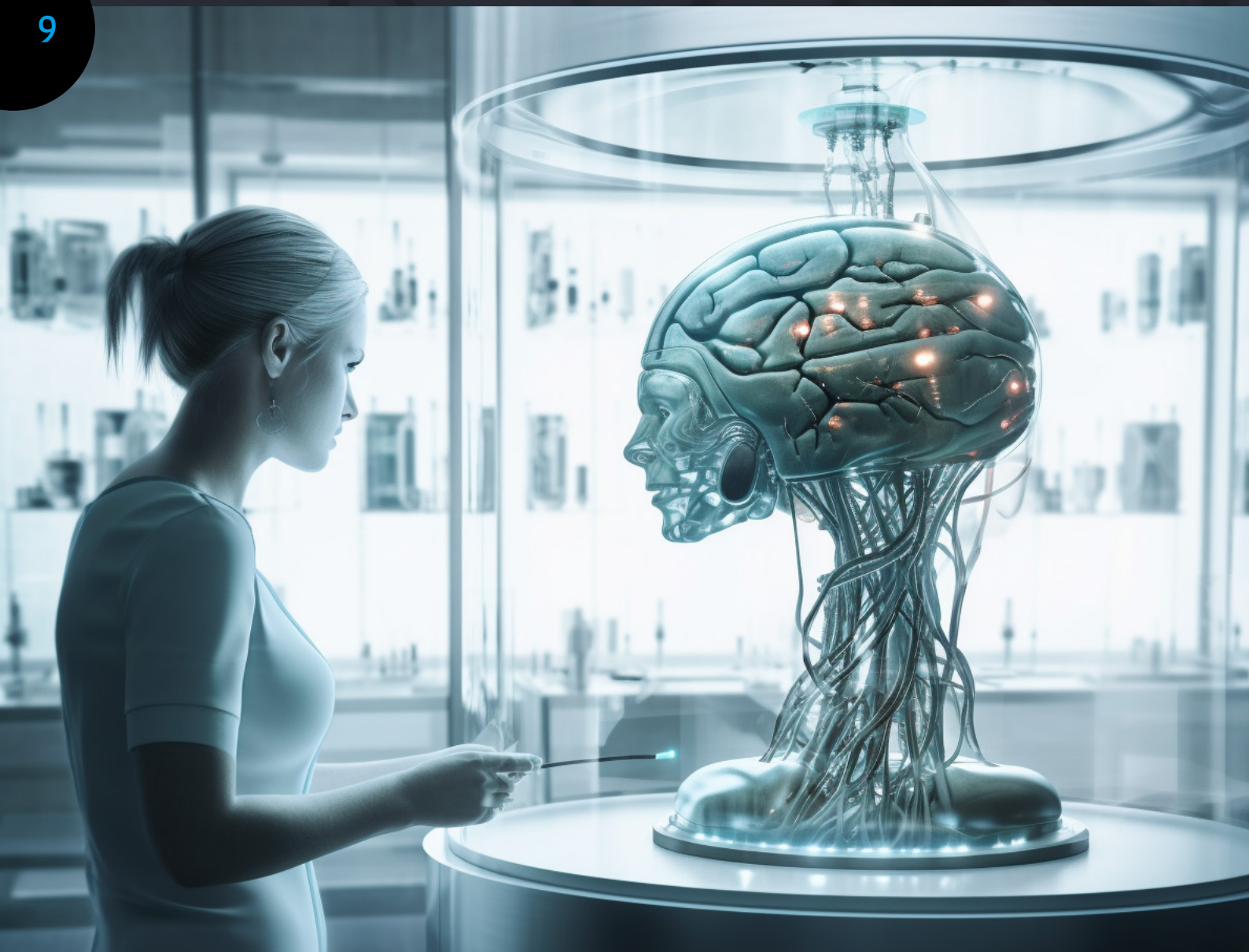
Moderní neuronové sítě se skládají z ohromujícího počtu neuronů, parametrů a vazeb mezi neurony. Vrstev těchto sítí (vrstva neuronů, které přijímají signály z předchozí vrstvy a posílají je do další) je více, hovoří se proto o hlubokých neuronových sítích. Právě tyto sítě jsou v pozadí současných mimořádných úspěchů na poli umělé inteligence, jako jsou sítě generující obrázky podle slovního zadání nebo velké jazykové modely jako je GPT-4. V průběhu učení se neustále mění parametry neuronů a jejich vstupů, které rozhodují o chování celé neuronové sítě. Učení neuronových sítí je založeno na optimalizaci tzv. účelové funkce neuronové sítě. Tato účelová funkce může být navržena např. tak, že její hodnota je velmi vysoká v případě, že neuronová síť je správně navržena pro řešení dané úlohy, a velmi nízká, případně velmi záporná, je-li neuronová síť navržena špatně. V počátku učení se vždy definuje nějaký počáteční stav neuronové sítě (který je skoro jistě nevyhovující pro řešenou úlohu) a z tohoto počátečního stavu se přechází do naučeného stavu (kdy síť poskytuje uspokojivé odezvy na vstupy) aplikováním běžných algoritmů numerické optimalizace (přes parametry neuronové sítě, tedy váhy a prahy neuronů). Jakmile je síť dostatečně natrénovaná, tyto parametry se zafixují a síť může sloužit účelu, pro který byla naučena.

JEDNODUCHÁ NEURONOVÁ SÍŤ

Síť obsahuje vstupní vrstvu neuronů, tzv. skrytou (druhou) vrstvu neuronů a konečně poslední, výstupní vrstvu neuronů.



V posledních měsících vzbuzují velký zájem (a nemalé obavy) komplexní neuronové sítě, jako je GPT-4. Jedná se o velký jazykový model, založený na neuronové síti o desítkách vrstev a stěží uvěřitelných 170 biliónů parametrů. Učí se na pro nás nepředstavitelně velkých souborech textů a využívá velmi sofistikované statistické nástroje, které jí umožňují odhalovat vazby mezi slovy, větami a kontexty. Díky tomu „rozumí“ konverzaci v přirozeném jazyce, dokáže plnit celou řadu příkazů, psát souvislé texty, překládat do mnoha jazyků, včetně mrtvých (latina, starořečtina), sumarizovat texty, provádět matematické operace, programovat a mnohé další. Je třeba si ale uvědomit, že tato a podobné sítě nemají žádné mentální stavy, nerozumí výrazům, které skládají dohromady, nemají úmysly, nemohou něco chtít, zamýšlet, o něco se snažit apod. Umějí dělat pouze to, na co byly vytvořeny a natrénovány. Nezapomínejme také, že fyzicky neexistují, na rozdíl od neuronů v biologickém mozku. Ve své podstatě jsou to strukturovaná data uložená v paměti počítače, která procesor načítá, zpracovává podle určitých instrukcí (zakódovaných do podoby „neuronů“) a ukládá zpět do paměti. Když „vypneme“ lidskou mysl (třeba během spánku), neurony dále existují a jsou velmi aktivní (i když si to neuvědomujeme). Pokud ale vypneme počítač, zůstane neuronová síť zachována jen v podobě fyzicky zapsaných dat (jako to platí pro všechna data v paměti), která se neliší od jiných typů dat (obrázků, textů, muziky) a nijak aktivní není.



Je umělá inteligence existenčním rizikem?

Dalo by se říct, že reflexi společenských a etických implikací AI do jisté míry rozděluje časové hledisko. Na jedné straně jsou autoři, kteří se zabývají současným využitím AI nebo blízkou budoucností, a spolu s tím i aktuálními problémy, které v jednotlivých oblastech potřebujeme vyřešit. Další skupinu pak tvoří lidé, kteří se zaměřují na výhledy do vzdálenější budoucnosti, ať už spojené s nereálnými očekáváními technooptimismu nebo katastrofickými hrozbami technopesimismu.

rámeček č. 3

CO JSOU TO EXISTENČNÍ RIZIKA?

Rizika můžeme chápat jako reálné možnosti, že nastane nějaká špatná událost. Můžeme je blíže vymezit prostřednictvím tří proměnných: rozsahu, intenzity a pravděpodobnosti. Rozsah vyjadřuje, jak velkou část lidské populace by případná událost zasáhla, intenzita určuje, jak špatná by událost byla, pravděpodobnost je vyjádřením šance, že událost nastane. První dvě proměnné umožňují klasifikovat rizika podle rozsahu na osobní (týkají se jen jednoho člověka), lokální, globální a dokonce transgenerační, podle intenzity na mírná, snesitelná a terminální (působící smrt či velmi drastické snížení kvality života). Za globální katastrofická rizika označujeme rizika, která jsou globální či transgenerační a současně snesitelná či terminální. Zvláštním případem těchto rizik, o němž se v součas-

nosti hovoří v souvislostí s umělou inteligencí, jsou existenční rizika. Kdyby nastala nějaká událost chápaná jako tento druh rizika, došlo by k vyhubení lidského života nebo takovému úpadku lidstva, že by se kvalita života dramaticky snížila a způsob naší existence by byl velmi vzdálený od toho, co si představujeme jako lidskou kulturu a civilizaci.

Vnímání umělé inteligence jako existenčního rizika (viz Rámeček 3) spadá právě do této druhé kategorie. Tyto představy a pochmurné predikce jsou spojeny s možností vzniku obecné umělé inteligence a superinteligence (viz Rámeček 4). Zatím se však jedná o pouhé spekulace. Základní argumentační linie technoskeptiků má zhruba následující podobu:

1. Jednoho dne vytvoříme obecnou umělou inteligenci.
2. Jakmile ji vytvoříme, pověříme ji úkolem vlastního vylepšování.
3. Zřejmě krátce na to vznikne superinteligence.
4. Superinteligence se bude dále vylepšovat.
5. Nastane exploze superinteligence a vznik stále inteligentnějších strojů, jejichž inteligence bude radikálně převyšovat inteligenci lidskou a zcela se vzdálí možnostech lidského pochopení a kontroly.

Zřejmě nejproslulejším zastáncem tvrzení o AI jako existenčním riziku je filosof Nick Bostrom působící na Oxfordské univerzitě, který se proslavil publikací Superintelligence (v češtině Superinteligence. Až budou stroje chytřejší než lidé. Praha: Prostor, 2017). K němu se přidávají někteří další známí odborníci jako Stuart Russell, Oren Etzioni, Max Tegmark, autoři popularizačních knih o existenčních rizicích Toby Ord, Juval Noach Harari, nebo známé osobnosti jako Stephen Hawking, Elon Musk, či Bill Gates.

rámeček č. 4

TYPY UMĚLÉ INTELIGENCE

Umělá inteligence se podle šíře svých schopností rozděluje na úzkou, obecnou a superinteligence. Úzká umělá inteligence dokáže řešit pouze úzce vymezenou skupinu problémů a často zde dosahuje mnohem lepších výkonů než lidé (šachy, go, rozpoznávání vzorů apod.). Obecnou umělou inteligenci můžeme chápat jako svatý grál současného výzkumu, jednalo by se o takový systém umělé inteligence, který by dokázal řešit stejnou škálu kognitivních problémů jako člověk. Konečně superinteligencí se rozumí inteligence (nejen umělá), která ve všech aspektech přesahuje lidské schopnosti. Někdy se také setkáváme s rozlišením mezi silnou a slabou umělou inteligencí. Silná umělá inteligence

je inteligence s lidskými vlastnostmi a schopnostmi, včetně sebeuvědomování a sebereflexe, zatímco slabá umělá inteligence tyto typicky lidské vlastnosti postrádá. Obecná umělá inteligence ani superinteligence nemusí nutně být silnou umělou inteligencí. Všechny současné systémy umělé inteligence představují úzkou a slabou umělou inteligenci.

Umělá inteligence v pesimistických scénářích některých autorů nemusí být nijak záludná a toužit vyhubit lidstvo. Strojové superinteligenci však jednoduše nebudeme rozumět, její architektura, fungování a způsoby řešení problémů budou zcela mimo dosah lidského chápání. Pověříme-li ji sebezdokonalováním, budou vznikat další, a ještě inteligentnější stroje, jejichž vývoj nebude pod lidskou kontrolou. Tyto stroje nemusí projevovat nepřátelské úmysly, jejich pokročilé inteligenci však mohou lidé připadat podobně, jako nám připadají mravenci: živí, ale nikoli tak důležití, aby stáli v cestě naplňování našich vlastních cílů. Superinteligence by dokonce mohla plnit úlohy, které jí zadal člověk, mohla by je však řešit takovým způsobem, který ohrozí naši existenci. A bude-li naplňování svých cílů považovat za důležité, zřejmě si vyvine obranné mechanismy, které nám zabrání ji jednoduše vypnout. Opět, není nutné si představovat, že bude mít něco jako pud sebezáchovy; prostě jen bude budovat mechanismy, které jí zajistí bezproblémové a co nejefektivnější plnění jejích cílů. Výsledek je ale pokaždé stejný: vyhubení lidstva, nebo jeho zotročení či třeba odsunutí na vedlejší kolej, což bude mít za následek ztrátu vlády nad stroji a postupnou degradaci na úroveň, v níž nebude existovat nic specificky lidského, jako např. kultura, respekt k lidské důstojnosti, autonomii, právům a svobodě.

Jsou tyto obavy reálné?

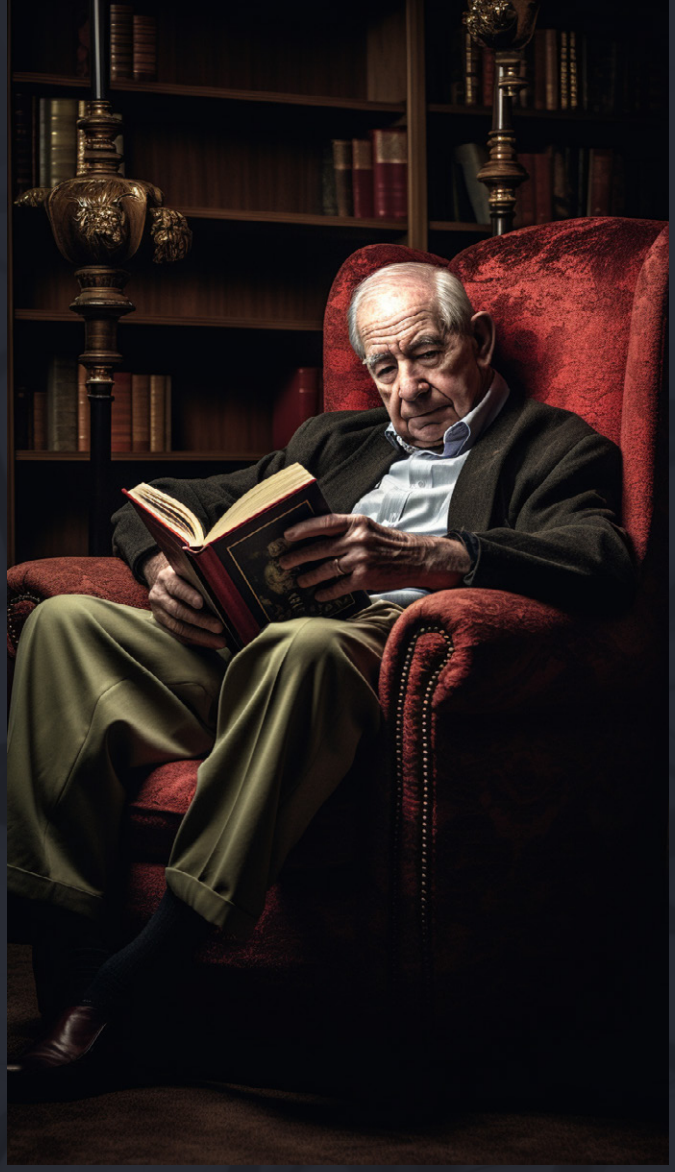
Umělá inteligence sama o sobě není takovým rizikem, před kterým nás někteří autoři varují. Je možné, že jednoho dne vytvoříme obecnou umělou inteligenci a je také možné, že se tím otevře cesta k superinteligenci. Zatím se ale k obecné umělé inteligenci ani zdaleka neblížíme. Je pravda, že současné neuronové sítě dokáží pozoruhodné věci a manifestují různé formy inteligentního chování, které již překračují lidské schopnosti v konkrétních oblastech. Musíme si ale uvědomit, že například velké jazykové modely, které jsou v posledních měsících středem pozornosti, vykazují inteligentní chování, aniž by ale měly vlastnosti, které si spojujeme s inteligencí lidskou. Dokážou provádět velmi sofistikované statistické operace na ohromujících množstvích dat, nemají ale žádné pochopení či porozumění tomu, co dělají. Neznají významy slov a vět, doopravdy nám nerozumí, nevědí, co nám odpovídají (vlastně ani neví, že odpovídají).

Americký filosof John Searle přišel s myšlenkovým experimentem, který je znám jako argument čínského pokoje. Představme si, že v jedné místnosti jsou všechny potřebné čínské znaky, které lze použít pro skládání vět. Nachází se tam také člověk, který čínsky neumí. Má ale k dispozici velmi podrobný manuál, který mu říká, jakým sekvencím slov v čínském jazyce (dotazům) přiřadit odpovídající sady čínských znaků (odpovědi). Když mu jedním okénkem předložíme dotaz napsaný

na papíře, dokáže v manuálu najít sekvenci čínských znaků, které jsou adekvátní odpovědí. Z vnějšího pohledu to vypadá, jako kdyby uměl čínsky, když ale prohlédneme mechanismus vzniku odpovědí, uvědomíme si, že tomu tak není. Searle chtěl tímto myšlenkovým experimentem prokázat, že počítače nikdy nebudou disponovat silnou umělou inteligencí (viz Rámeček 4), protože provádějí pouze syntaktické operace (operace se znaky) a postrádají jejich pochopení.

GPT-4 je vlastně takovým obrovským čínským pokojem, s jedním důležitým rozdílem. Nedostal předem k dispozici manuál, ale musel si ho vytvořit sám tím, že prozkoumával a učil se statistické vztahy mezi slovy, frázemi, větami a kontexty. Jinak se ale od čínského pokoje neliší: když dostane dotaz, identifikuje jeho části a ve shodě se svým manuálem mu přiřadí odpovídající výstup, který si jako odpověď můžeme přečíst. V pozadí toho všeho jsou obrovská data, mimořádný výpočetní výkon moderních počítačů, neuronové sítě s desítkami vrstev a bilióny parametrů a jejich nastavení prostřednictvím učení.

Umělá inteligence, alespoň v současné podobě, existenčním rizikem není. Skutečná podoba obecné umělé inteligence, která by mohla stát na začátku cesty k realizaci existenčních rizik, nám ve skutečnosti není příliš známá a technopesimisté se svými úvahami pohybují v oblastech velkých neznámých. Je docela dobře možné, že namísto nějakého stroje s obecnou umělou inteligencí budou vznikat systémy, které v některých oblastech budou mít inteligenci nulovou (budou například postrádat širokou škálu lidské praktické inteligence), v jiných lidskou a v dalších zase nadlidskou. Půjde ale o spojování úzkých inteligencí plně limitovaných svou architekturou, jejíž omezení nebudou schopné překročit. V každém případě si ale myslíme, že úvahy o rizicích spojených s obecnou umělou inteligencí jsou předčasné a odvádějí naši pozornost od problémů, které jsou až příliš reálné. Všechny tyto problémy spojuje umělá inteligence, nikoli však v roli zlověstného aktéra, ale spíše nástroje, který my lidé používáme špatně a pro morálně nesprávné účely.



John Searle

V následující sekci představíme několik aktuálních problémů, které poukazují na to, že umělé systémy s inteligentním chováním nevyvíjíme, nenavrhujeme a nevyužíváme rozumně, moudře a morálně správně. Zmíníme také některé výzvy, které umělá inteligence vyvolává v oblasti práva. Mějme ale na paměti, že rozlišení těchto výzev na etické a právní může být v některých případech poněkud umělé a oba normativní systémy regulace lidského chování by měly při hledání jejich řešení postupovat ruku v ruce.

Některé aktuální výzvy

#ETIKA

Neuronové sítě, o které se většina současných systémů AI opírá, představují nesmírně komplikované struktury, které u každého druhu učení vyžadují obrovské množství opakování základních kroků. Síti se předloží vstup, ta jej zpracuje, úspěšnost zpracování se vyhodnocuje a následně se formou zesilování nebo zeslabování spojení mezi vnitřními prvky sítě vytvářejí nová nastavení. Tyto kroky se mnohokrát opakují; a protože se síť s každým krokem mírně pozmění, není divu, že brzy získá konfiguraci, která je i pro jejího autora nebo lidského pozorovatele nepochopitelná. Tomuto jevu se říká problém černé skříňky (blackboxing); sítě dělají svou práci mimořádně dobře, nevíme ale dostatečně dobře jak.

Problém černé skříňky má několik důsledků, které souvisí s tím, že nám rozhodovací procedury, jimiž dospívají k výsledkům, částečně zůstávají skryté. Když AI rozhoduje o tom, zda máme dostat hypotéku, být povýšeni v zaměstnání, nebo pokáráni či dokonce propuštěni, když predikuje naše chování, například pravděpodobnost spáchání dalšího zločinu apod., ztrácíme možnost tato rozhodnutí pochopit, zhodnotit, podrobit kritice či se vůči nim odvolat. Vynořují se proto snahy o vývoj vysvětlitelné umělé inteligence XAI (X je zkratka pro explainable), která by odhalovala způsoby, jimiž produkuje své výstupy. Zůstává ale zatím otevřenou otázkou, zda snaha o vysvětlitelnost nesnižuje efektivitu AI a do jaké míry je dostatečné vysvětlení skutečně možné.

Další závažný problém představují vnitřní rozhodovací algoritmy neuronových sítí. Jedná se v podstatě o optimalizační algoritmy (hledající nejlepší řešení, jak spojit vstupy (např. fotografie lidských obličejů) s výstupy (např. kategorizace kriminálník/nevinný člověk)) a je vysoce nepravděpodobné, že by k tomu využívala lidsky uchopitelné kategorie. Ukazuje se, že si sítě ze vstupů vytvářejí pro nás nepochopitelné vzory, které z lidského hlediska často vůbec nedávají smysl. To nemusí být vždy problém: asi nám může být celkem jedno, jakým postupem se AI dopracovala k tomu, že nám na základě našich preferencí doporučila ten nejlepší parfém. Když ale jde o doporučení, které jsou citlivá na lidsky uchopitelné a třeba i hodnotově podbarvené kategorie, dostáváme se do potenciálního konfliktu. Například doporučení, jak se chovat k druhému člověku, závisí na mnoha pro nás samozřejmých kategoriích jako je věk, pohlaví, mocenské postavení, zranitelnost, rasa apod. Neuronové sítě ale nechápou pro nás zřejmý význam těchto kategorií a neumí je adekvátně začlenit do svých rozhodovacích procesů. Zcela legitimní je proto otázka, zda bychom se jejími doporučeními v těchto citlivých oblastech měli řídit. Například systém KOMPAS, který s jistou mírou úspěšnosti předpovídá recidivu, je v individuální rovině nespravedlivý vůči Afroameričanům. Chyba ale není na straně AI, nýbrž lidí, kteří ho navzdory tomuto faktu i nadále používají.

Moderní umělá inteligence je závislá na velkých datech, která jí umožňují se efektivně učit (nastavovat vnitřní parametry neuronových sítí) a plnit úkoly. Kvalita výstupů je ale zásadně závislá na kvalitě těchto dat. Jsou-li například v trénovacích sadách dat nedostatečně zastoupeny menšiny, bude v jejich případě AI poskytovat méně kvalitní odpovědi. Může dokon-

ce diskriminovat či neférově distribuovat rizika. Například autonomní vozidla mohou mít problémy dostatečně dobře rozpoznat lidi jiné barvy pleti, než je bílá, proto jim hrozí větší rizika újmy. Je-li sada dat zvolena nevhodně, může AI snadno převzít celou řadu předsudků, které jsou bohužel ve společnosti stále přítomné. Chyba ale opět není na straně AI. Pokud se například nějaký chat založený na strojovém učení „stane“ rasistou, je to jen proto, že rasisty jsme my lidé a trénovali jsme ho na reálných konverzacích, které jsou bohužel rasismu, sexismu a dalších nemorálních postojů plné. Musíme proto převzít zodpovědnost za výběr dat a zajistit, aby reprezentovala všechny lidi rovně, aby nezahrnovala naše předsudky, nemorální postoje a kognitivní zkreslení.

Velké obavy vzbuzuje fakt, že mnohé lidské aktivity jsou v menší či větší (či dokonce úplné) míře automatizovatelné. Tento trend se s pokroky na poli AI zrychluje a mnohé pracovní pozice jsou tak v ohrožení. Řada novinových titulků varuje, že nám „stroje vezmou práci“, ale to je omyl. Lidi nahrazují stroji jiní lidé, umělá inteligence je v tom nevinně. Neměli bychom přenášet odpovědnost na ni, protože je pouze naše. Máme k dispozici všechny nástroje, které mohou zajistit přerozdělování bohatství generované stroji všem lidem, zvláště těm potřebným, aby chudoba nenarůstala ale naopak klesala. Lidskou práci také nelze redukovat na činnosti, které vykonáváme jen proto, abychom si vydělali na živobytí. Zajištěný nepodmíněný příjem v odpovídající výši nemusí znamenat, že lidé „bez práce“ složí ruce do klína; právě naopak. Mohou se začít věnovat aktivitám, které dávají jejich životům smysl a přispívat k dobrému životu celé společnosti. Nesmíme však být pasivní a jen zpětně řešit nastupující problémy. Být aktivní neznamená řešit to, co se už stalo, ale zaměřit se na budoucnost, převzít za ni odpovědnost a nastavovat podmínky společné existence ve společnosti takovým

způsobem, aby umožňovala kvalitní a dobrý život všem jejím členům.

Další vážný problém, který AI nevytvořila, ale významně posílila, se týká našeho poznání. V posledních letech se začíná hovořit o pandemii špatného myšlení. Lidé všech vrstev nedokážou kriticky myslet, správně vyhodnocovat informace a ověřovat si jejich zdroje. Z aktivních uživatelů nástrojů lidského rozumu se proměňujeme na pasivní recipienty všeho možného, co nám média a algoritmy předkládají, bez ohledu na věrohodnost a pravdivost. Usnadňuje nám to uzavírat se v kamerách ozvěn a informačních bublinách, kde se navzájem podporujeme v často velmi škodlivých názorech realitě na hony vzdálené. Ale ani za to bychom umělé inteligenci vinu dávat neměli. Nenutí nás myslet špatně a podléhat dezinformacím či propagandě. Na vině je opět naše pasivita a neochota kultivovat nástroje kritického myšlení, které by nám umožnily dobrou orientaci ve stále komplexnějším a rychle se vyvíjejícím světě.

#PRÁVO

Z právního pohledu musíme čelit rizikům, které může vývoj a použití AI přinášet. Těmi nejvýznamnějšími jsou:

1. Zásahy do soukromí.
2. Zneužívání osobních údajů a dat obecně.
3. Zkreslení (bias) a předsudky vedoucí k diskriminaci.
4. Odpovědnost za újmu způsobenou systémy AI.

Pro některé situace najdeme řešení již v současném právu, zatímco jiné vyžadují novou regulaci.

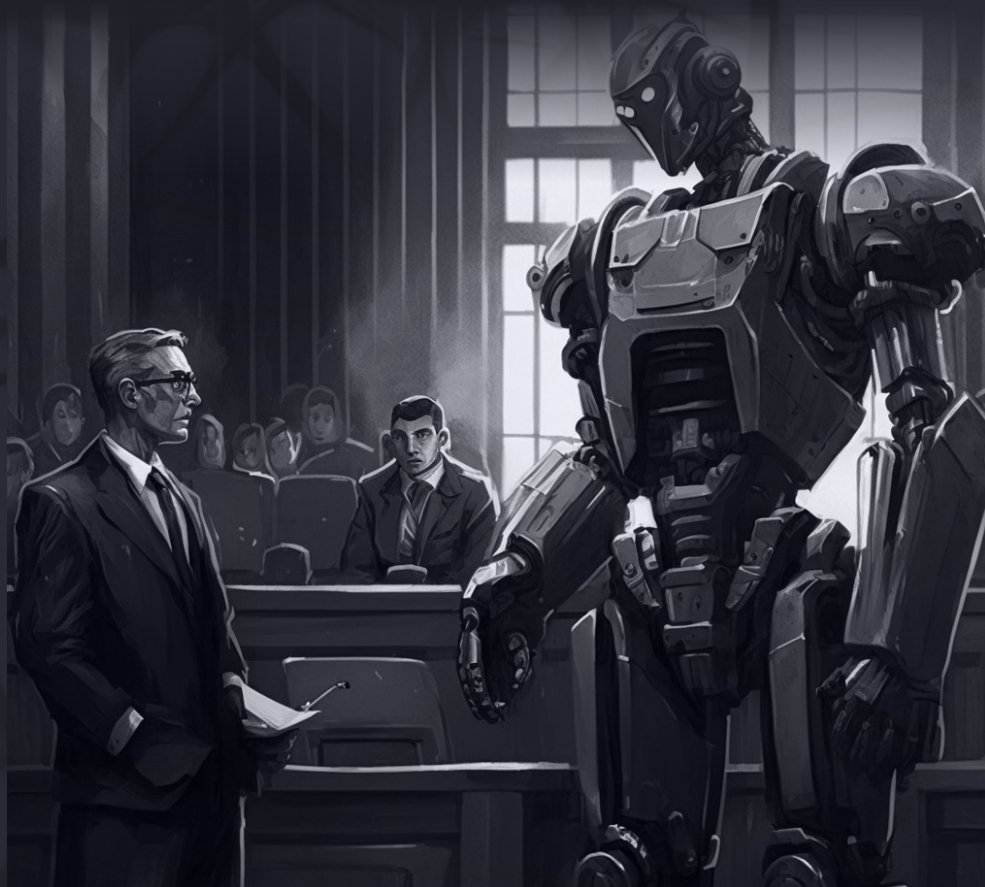
Kvalitní právní úprava by měla hledat rovnováhu. Mezi individuálními právy a ochranou společnosti před nastíněnými riziky na jedné straně a snahou zasahovat co nejméně s cílem nesnižovat konkurenceschopnost aktérů, kteří AI vyvíjejí či s ní jen pracují.

V souvislosti s rozvojem a popularizací především velkých jazykových modelů se plány na regulatorní rámce nyní diskutují celosvětově. Na evropské úrovni se aktuálně schvaluje několik předpisů v čele s AI Nařízením (AI Act). Nařízení pracuje s rozdělením AI systémů do několika stupňů podle míry rizikovosti. Návrh je však opakovaně terčem kritiky, a to často velmi oprávněně.

Hned prvním úskalím, na které předpisy řešící AI naráží, je samotná její kvalitní právní definice. Aby stručně a srozumitelně zahrnula vše podstatné, nebyla ani příliš vágní ani příliš úzce vymezená. Na dobré definici stojí a padá celá aplikace právní normy.

Celounijní regulaci prozatím předbíhají některé dílčí národní úpravy či opatření formou zákazů. I u nich se však projevuje mediální zkratkovitost, nepodložené extrémní nadšení i naopak skepse a pudový strach z neznámého. Všechny tyto jevy mohou vést k tomu, že se na AI nahlíží z nevhodných, omezených, čistě lidských úhlů pohledu.

Jestliže je umělá inteligence komplexní oborem s ambicí postihnout inteligenci v celé její šíři, minimálně obdobnou té lidské, pak stejně komplexní musí být i její právní rámec. Kromě čistě právního pohledů musí reflektovat i etické, sociologické, psychologické a samozřejmě technické aspekty.



Od pasivity k aktivitě

Technologie nás stále více ovlivňují a pronikají do každého zákoutí našeho života, života jednotlivců, skupin i celé společnosti. Mnozí lidé se stávají bezradnými vůči již existujícím technologiím a algoritmům, které je ovládají. Někteří toto neznámé vítají a mají od něho nereálná očekávání, zatímco jiní se ho obávají a mají k němu odpor. V obou případech se jedná o postoje, které nejsou založeny na porozumění realitě, a proto mají řadu negativních důsledků. Důležitou součástí dobrého a kvalitního života v moderní společnosti je proto vhled do povahy, fungování, přínosů a také limitů jednotlivých technologií, zejména v oblasti bouřlivě se rozvíjející AI. Reálné zhodnocení je možné pouze v případě, že se vyhneme naivnímu technooptimismu, podle kterého jsou veškeré problémy řešitelné pomocí technologií, a stejně tak technopesimismu, podle kterého jsou technologie kořenem všech problémů, jimž čelíme.

Snažili jsme se upozornit na to, že hlavním problémem není samotná umělá inteligence, ale lidská pasivita a neochota přijmout plnou odpovědnost za její vývoj a dobré využívání v lidské společnosti, které bude vedeno snahou prospět každému z nás. Užívání nástrojů s inteligentním chováním by mělo být rozumné, moudré a morální.

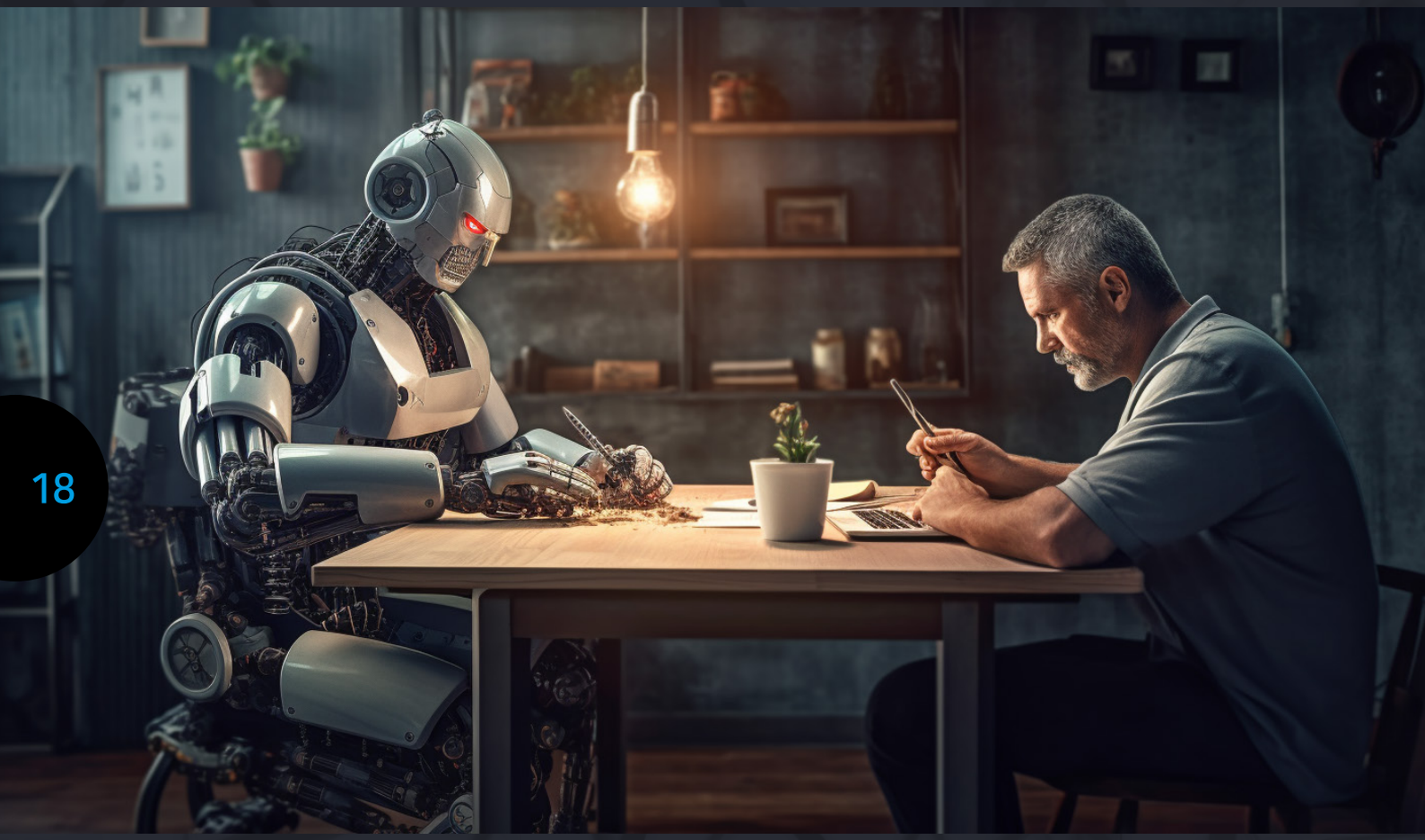
rámeček č. 5

ALGORITMICKÁ GRAMOTNOST

Algoritmickou gramotností rozumíme základní teoretické a praktické znalosti a dovednosti, které nám umožní využívat nástroje umělé inteligence rozumně (s ohledem na jejich možnosti), moudře (s vědomím si jejich omezení a nutnosti kritického myšlení při jejich využívání) a morálně správně (ve prospěch každého z nás a celé společnosti).

Mělo by být rozumné, to znamená, že se musíme učit využívat nástrojů umělé inteligence způsoby, které jsou v souladu s jejím určením a umožní nám co nejefektivnější řešení celé plejády úloh. Mělo by být moudré, což znamená, že si budeme dobře vědomi omezení AI. S tím je spojeno, že její výsledky nebudeme přijímat pasivně, nekriticky a budeme umět rozlišovat situace a aplikační oblasti, v nichž lze AI důvěřovat více od těch, kdy je moudré důvěřovat jí méně. A konečně by mělo být morální. Znamená to minimálně dvě věci. V první řadě to, že nástroje umělé inteligence budeme využívat tak, aby byl zachován respekt k lidské důstojnosti, základním lidským hodnotám a právům, férovosti a spravedlivé distribuci rizik a benefitů. A v druhé, neméně důležité řadě potom to, že se návrh a vývoj systémů umělé inteligence bude ve všech fázích snažit začlenit do rozhodovacích mechanismů systémů umělé inteligence respekt k lidským právům a hodnotám.

Umělou inteligenci nemusíme vnímat jako riziko, jako jakéhosi cizorodého aktéra, který se nutně vzepře svým stvořitelům. Tento dualistický (my – stroje) antagonismus je nesprávný. Systémy umělé inteligence jsou mimořádnými nástroji, které bychom měli vnímat jako fyzické a kognitivní rozšíření našich specificky lidských vlastností, včetně inteligence, rozumnosti a moudrosti. Namísto soupeření se snažme o koevoluci. Lidé a stroje se mohou společně vyvíjet tak, aby naše individuální životy a společnost vzkvétaly. Všechny nástroje, které potřebujeme (včetně etiky a práva) máme k dispozici a je jen na nás, abychom se probrali z dogmatického spánku a poraženecké strnulosti a začali aktivně jednat.



AUTORSKÝ KOLEKTIV

David Černý (filosofie e etika AI): david.cerny@cs.cas.cz
František Hakl (neuronové sítě, strojové učení): hakl@cs.cas.cz
Tomáš Hříbek (filosofie vědomí, AI): tomas_hribek@hotmail.com
Juraj Hvorecký (filosofie a etika AI): juraj@hvorecky.com
Linda Kolaříková (právo AI): linda.cechova@gmail.com
Monika Mareková (právo AI): monika.marekova@gmail.com
Daniel Novotný (filosofie, etika AI): novotnyd@tf.jcu.cz
Emil Pelikán (strojové učení, AI): pelikan@cs.cas.cz
David Procházka (kritické myšlení, AI): david.prochazka.km@vse.cz
Michal Trčka (etika, nanotechnologie): mich.t@centrum.cz
Jiří Wiedermann (AI, matematika, informatika): jiri.wiedermann@cs.cas.cz

Akademie věd České republiky

STRATEGIE AV21

Špičkový výzkum ve veřejném zájmu

VŠECHNY OBRÁZKY V TÉTO PUBLIKACI BYLY VYGENEROVÁNY SYSTÉMEM MIDJOURNEY sazba Alena Kučerová, 2023

ISBN: 978-80-87136-23-2