Jak přemítat o umělé inteligenci*

Jiří Wiedermann

Centrum Karla Čapka pro výzkum hodnot ve vědě a technice
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha, Česká republika
Email: jiri.wiedermann@cs.cas.cz

Motto: *„Předpokládám, že moudrost znamená správné používání znalostí. Samotné vědění ještě neznamená moudrost. Je mnoho lidí, jež vědí mnoho — ale to z nich dělá ještě větší hlupáky. Není nad hlupáka než vědoucího hlupáka. Poznání, jak používat znalosti, znamená mít moudrost.“*

C. H. Spurgeon, *The Fourfold Treasure* (1871)

Abstrakt

„Proč rozvíjíme umělou inteligenci?“, a *„Jaký účel má používání umělé inteligence?“* Příhodnou odpověď naznačují již současné velké jazykové modely: umělou inteligenci rozvíjíme za účelem získání a aplikování umělé moudrosti, jež umožní dělat moudrá rozhodnutí a chovat se moudře. Ukážeme, že současné jazykové modely nepřímo, vzhledem ke svým jazykovým schopnostem a návaznosti dat, ze kterých se učí, na popisy z reálného světa, dovedou extrahovat sémantiku ze syntaktických dat a navíc splňují podmínky tzv. 4E kognice (embodied, embedded, extended, enacted cognition). Tato se snaží vysvětlit mechanismy inteligentního chování pomocí dalších než výlučně výpočetních prostředků. Tyto modely tak vládnou jistou formou iluzorní inteligence a iluzorní moudrosti, doposud nepopsanou v odborné literatuře. Toto poznání má fundamentální význam pro filozofii a metodologii výzkumu umělé inteligence, protože představuje zásadní posun ve výpočetním paradigmatu používaném v umělé inteligenci, a sice od pohledu na umělou inteligenci jako na procesy generující znalosti směrem k procesům generujícím a využívajícím moudrost.

1 Úvod

Hledáme odpovědi na dvě zdánlivě jednoduché otázky: *„Proč rozvíjíme umělou inteligenci?“,* a *„Jaký účel má používání umělé inteligence?“* Samozřejmě, zajímají nás netriviální odpovědi, vycházející z hlubšího pochopení pojmu umělé inteligence, vyplývají z nějaké teorie, jež platí pro „jakoukoli umělou inteligenci“, přinesou nové vhledy do povahy umělé inteligence a dovolí extrapolaci trendů v oblasti umělé inteligence. Pod pojmem „jakákoliv umělá inteligence“ rozumíme jak to, co v současné době považujeme za umělou inteligenci, tak i veškeré druhy uměle vytvořené inteligence v budoucnosti, jak na Zemi, tak kdekoliv ve Vesmíru.

Nabízíme následující odpovědi: umělou inteligenci rozvíjíme za účelem získání a aplikování umělé moudrosti, jež umožní dělat moudrá rozhodnutí a chovat se moudře ve světě, ve kterém umělá inteligence operuje a má o něm dostatečné znalosti. V tomto kontextu chápeme umělou moudrost jako „správné používání znalostí“ pomocí účelného chování — kombinovaný efekt kognice a akce (viz úvodní citát). Nositeli umělé moudrosti jsou autonomní vtělení kognitivní behaviorální agenti. Ukážeme, že již v současné době se umělá moudrost vyskytuje v různých druzích a intenzitě a v rozličných světech, ve formě různých velkých jazykových modelech.

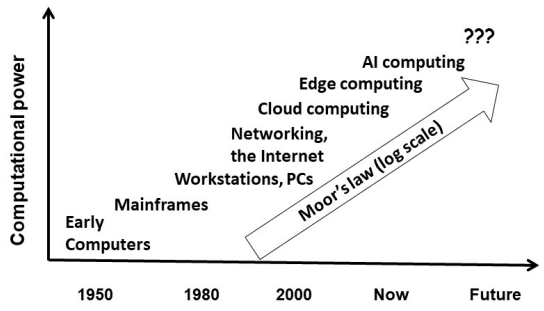
Pojmem moudrosti se zabývali především filozofové. Koncept moudrosti se vyskytuje v oblasti umělé inteligenci již od samotných počátků této discipliny. Nicméně, až donedávna bylo více pozornosti věnováno znalostem nežli moudrosti. Bylo tomu tak pravděpodobně proto, že znalosti jsou považovány za základní koncept, od nějž jsou odvozovány další koncepty, a speciálně moudrost. V posledních letech se situace mění a moudrost se dostává do popředí zájmu v oblasti praktické filozofie a etiky — a to je vlastně i náš případ.

V tomto kontextu práce přináší několik důležitých výsledků a poznatků.

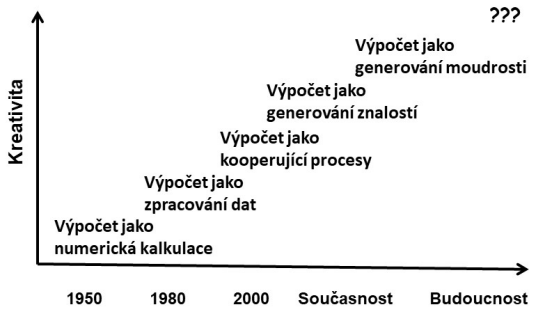
Po prvé, ukazuje, že systémy generující moudrost jsou přirozeným pokračováním trendů ve vývoji jak výpočetních technologií, tak i jejích aplikací na oblasti, vyžadující kognitivní schopnosti v celém spektru od úrovně získávání dat až po úroveň jejich zpracování a využívání.

Po druhé, zabývá se samotnou definici moudrosti, se kterou by mohly pracovat výpočetní technologie. Za tím účelem definuje model vtělených kognitivních behaviorálních agentů, jako systémů umělé inteligence, jež generují moudrost v širokém spektru různých světů. Tímto způsobem zobecňují pojem moudrosti i na domény odlišné od těch, ve kterých vy-

*Tato práce vznikla na základě diskusí a společných publikací s Janem van Leeuwenem z Utrechtské university.



Obr. 1. Vývoj počítáni



Obr. 2. Výpočetní trendy

niká lidská inteligence.

Po třetí, zkoumá současné velké jazykové modely z hlediska jejich využití pro generování moudrosti. Dochází k zajímavým výsledkům, že tyto systémy částečně splňují podmínky tzv. 4E kognice (embodied, embedded, extended, enacted cognition), která se snaží vysvětlit mechanismy inteligentního chování i dalšími než výlučně výpočetními prostředky. Jedná se o nepřímou, zprostředkovanou kognici, vyznačující se jakýmsi off-line povědomím o virtuálním světě, jež je zprostředkovaný prostřednictvím schopnosti modelu v jistém smyslu porozumět jazyku. Je to jakási zatím v odborné literatuře nepopsaná forma iluzorní inteligence, jež je podstatně odlišné od lidské inteligence a je charakteristická pro současné velké jazykové modely tím, že ukazuje možnost existence nějaké formy inteligence bez kognice.

Po čtvrté a v neposlední řadě, z filozofického hlediska a z pohledu metodologie, umělá moudrost představuje podstatný posun ve výpočetním paradigmatu používaném v umělé inteligenci, a sice od pohledu na výpočty jako procesy generující znalosti směrem k procesům generujícím a využívajícím moudrost. Koncept umělé moudrosti, vedle lidské moudrosti, se stává novou, a pravděpodobně finální metou lidského snažení.

Cílem práce není poskytnout návod, jak realizovat systémy generující moudrost, ale jak na ně pohlížet, jak je chápat a rozumět jejím možnostem a limitům.

2 Trendy ve využívání výpočetních technologií

Vývoj výpočetních technologií od jejich počátků v polovině 20. století až po současnost je obecně znám a přehledně jej zachycuje Obr. 1.

Z hlediska této práce je více zajímavý pohled na tyto technologie (Obr. 2), jež zachycuje jejich „informační sílu“ — pojem, jenž je zřejmější z popisu tohoto obrázku. Z obou obrázků je jasně vidět, jak s rozvojem informačních technologií a jejich aplikací rostla (a roste) jejich informační síla v rámci tzv. DIKW hierarchie: *data, informace, znalosti, moudrost* (Pyramid 2023). Tato hierarchie zachycuje skutečnost, ze v typickém případě je informace definovaná pomocí dat, znalosti pomocí informací, a moudrost pomocí znalostí.

V další kapitole ukážeme, jak lze na výše zmíněnou hierarchii pohlížet z hlediska teorie epistemických výpočtů, tj. z pohledu na výpočty jakožto procesů generujících znalosti.

3 Výpočet jako proces generující znalost

Základem našeho přístupu k výpočetním schopnostem systémů umělé inteligence je *epistemická teorie výpočtů* jež vychází z prací autorů van Leeuwena a Wiedermanna (2014, 2015a, 2015b, 2017). Z hlediska této teorie výpočtů nahlížíme na výpočty jako na procesy, které generují znalosti nad danou znalostní doménou $\mathbb{D}$ v rámci příslušné znalostní (epistemické) teorie $\mathcal{T}$. Výpočet pracuje tak, že kombinuje prvky znalostní domény — jimiž jsou *informace* (resp. jejich reprezentace), nazývané také *elementární znalosti* — do odvozených, často složitějších konstrukcí, které již tvoří *novou znalost*, opět nad danou doménou a v rámci teorie $\mathcal{T}$. Pro kombinaci těchto prvků používá výpočet množinu (odvozovacích) *pravidel,* která může být předem daná v rámci teorie $\mathcal{T}$, anebo se může tvořit pomocí učení během velkého počtu různých výpočtů nad danou doménou.

Systém tímto způsobem pracuje s více či méně formální teorii $\mathcal{T}$, která zachycuje vlastnosti dané znalostní domény a způsoby odvozování nových znalostí,

stále v rámci dané domény. Jakmile systém načte nějaká data, tyto se stávají v rámci teorie $\mathcal{T}$ informací. Z nich systém shora popsaným způsobem generuje znalosti; některé z nich se mohou stát výstupem výpočtu.

Díky své obecnosti znalostní přístup lze uplatnit nejen v dobře formalizovatelných, tzv. *exaktních znalostních doménách,* ale i ve znalostních doménách a pro odvozovací pravidla, které se vzpírají jakékoliv formalizaci. Takovým doménám budeme říkat *popisné znalostní domény.*

Příkladem formální znalostní domény budiž množina přirozených čísel s příslušnou teorii reprezentovanou pomocí Peanových axiomů.

Typickým případem popisné domény s neformálními odvozovacími pravidly je reálný svět. Jeho objekty, jevy, akce a vztahy mezi jimi jsou popsány pomocí přirozeného jazyka. Znalosti o takové doméně jsou zachyceny ve větách přirozeného jazyka. Odvozovací pravidla jsou v tomto případě tzv. *pravidla racionálního uvažování a chování.* Tato pravidla vycházejí z faktů a argumentů, která lze odpozorovat z přirozeného jazyka a zachytit v přirozeném jazyce. V typickém případě mají popisné domény rozsáhlé znalostní báze (jako např. obsah internetu) a relativně krátké odvozovací řetězce. Současné velké jazykové modely (LLM) jsou pěkným příkladem takových domén a neformálních teorií.

Pro formalizaci výše zmíněného přístupu viz práci van Leeuwena a Wiedermanna (2017). Přehled dosavadních výsledků z oblasti kognitivních výpočtů lze nalézt v práci Wiedermanna a van Leeuwena (2015a). Současně to umožní nový pohled na velké jazykové systémy, jak uvidíme v části 5.

V rámci epistemického přístupu k výpočtům lze tedy dobře popsat první tři úrovně DKIW hierarchie: data, informace, znalosti. Poslední, nejvyšší úroveň moudrosti popíšeme v další kapitole.

4 Od znalosti k moudrosti

Pojem moudrost, jako všechna slova přirozeného jazyka, která vznikly v běžném životě, je velmi kluzký, těžce definovatelný. Problém je v tom, že se jedná o slovo — kufr, jak to nazýval Marvin Minsky (1998). Jsou to slova, kterým lidé připisují, resp. do nichž „balí“, leckteré další významy. Např. Wikipedie definuje moudrost jako *vědění, mudrlanství, chytrost, slovo označující schopnost používat znalosti, pochopení, selský rozum a vhled* (Wisdom 2023). Co slovo, to další kufr. To není definice, ze které by se dalo vycházet ve výpočetním prostředí.

Naštěstí máme zde i definice z oblasti filozofie. Jedna z nich, použitelná v našem případu, je v úvodu této práce. Pro naše účely se hodí obecně akceptovaná „slovníková“ definice: *zatímco znalost je definována jako nabytí dat a informací, moudrost je praktická aplikace a použití znalostí za účelem vytváření hodnot.* Tato definice klade důraz na praktické aspekty moudrosti — její využití v reálném světě.

Pro naše další účely si tuto definici ještě upravíme. Všechny známe definice moudrosti se vztahují k lidské moudrosti, a vznikly, pochopitelně, na základě kognitivního (smyslového) poznání světa. My budeme potřebovat definici umělé moudrosti vhodnou pro použití v v obecných, a tím pádem také v umělých, kognitivních systémech. V obecném případě kognitivní systém je autonomní systém schopný vnímat své okolí, učit se ze své zkušenosti, předvídat běh událostí, jednat účelně a eticky pro splnění svých cílů a přizpůsobovat se měnícím okolnostem (Vernon 2021).

V tomto kontextu budeme používat následující definici: *přirozená i umělá moudrost je správné používání znalostí pomocí účelného chování — kombinovaný efekt kognice a jednání směřující k vytváření pragmatických či dodržování etických hodnot.*

Nositeli umělé moudrosti jsou především autonomní interaktivní vtělení kognitivní behaviorální agenti. Umělá moudrost se vyskytuje v různých světech, druzích a intenzitě závislé na ustrojení a schopnostech agenta získávat a využívat potřebné znalosti ve svém osvětí. V naší terminologii epistemických výpočtů různé světy (osvětí) reprezentuje znalostní doména $\mathbb{D}$, druh umělé moudrosti odvozujeme od epistemické teorie $\mathcal{T}$ — jaká je její vyjadřovací síla, co všechno, jaká fakta, vztahy a dotazy dovede vyjádřit, intenzita hovoří o efektivitě takového vyjádření, ustrojení se týká vybavení agenta pomocí senzorů a efektorů a repertoáru jejich akcí. Problémem ovšem zůstává část definice, požadující *„správné používání znalostí pomocí účelného chování“*, a také, co se chápe pod pojmem *„směřovat k vytvoření pragmatických či etických hodnot“*? Co to znamená v kontextu autonomních interaktivních vtělených kognitivních behaviorálních agentů?

To je složitá otázka, která sa týká množství a kvality znalostí agenta a jeho schopnosti je využívat.

Intuitivně, být moudrý v nějaké znalostní doméně znamená, že jednak agent má pomocí svých senzorů přístup ke všem objektům v této doméně, a také, že se dokáže vypořádat se všemi situacemi, dotazy a příkazy, týkající se těchto objektů, avšak pouze vzhledem ke svému poslání. Vše k tomu potřebné je popsáno v epistemické teorii. Poslední výhrada je důležitá — např. nemůžeme chtít po autonomním vozidle, aby s námi konverzovalo o všech objektech, které zachycuje svými senzory, avšak nejsou popsány v epistemické teorii, anebo aby se chovalo účelně v prostředí, pro které nebylo navrženo resp. naučeno — třeba v prostředí středověkého města.

Pro zodpovězení této výhrady musíme zavést pojem smysluplnosti (resp. poslání) agenta. Tento pojem je formálně popsán v jeho *funkční specifikaci* Φ. Tato specifikace popisuje, jaké funkce musí agent vykonávat a za jakých podmínek. Specifikace musí

splňovat následující dvě podmínky:

- Specifikace musí předepsat, jak se má systém chovat, v závislosti na teorii $\mathcal{T}$ v dané situaci s_i za předpokladu, že víme, jak se choval v předchozích situacích $s_1, s_2, \ldots, s_{i-1}$, pro libovolné $i > 1$ a libovolnou posloupnost situací, která se může v doméně $\mathbb{D}$ vyskytnout.
- Specifikace musí garantovat vytváření pragmatických hodnot a současně i dodržování etických hodnot.

Uvedená definice funkční specifikace autonomního interaktivní vtěleného kognitivního behaviorálního agenta je typická pro epistemický přístup, protože vyžaduje splnění dvou podmínek, aniž by něco hovořila o tom, jak toho dosáhnout, např. jestli musejí agenti být v nějakém smyslu inteligentní anebo ne. To ji dává širokou platnost — odpovídající systémy mohou být fixní, neměnné během své činnosti, a nebo se mohou učit, rozvíjet své znalosti (tzv. evoluční systémy), mohou být vědomé, mít svobodnou vůli a další mentální schopnosti. Pro specifikaci není důležité, jak má agent dosahovat svých cílů, ale je důležité to, co má dělat (vytvářet pragmatické hodnoty), a také, aby přitom dodržoval etické hodnoty či principy. V obecném případě jsou etické hodnoty popsány pomocí etické teorie $\mathcal{E}$. Je to opět specializovaný druh epistemické teorie, která popisuje etické principy — zásady a limity chování — které musí autonomní interaktivní vtělený kognitivní behaviorální agent dodržovat.

Teď konečně můžeme definovat umělou moudrost: *agent A se chová moudře, anebo stručně, je moudrý ve své doméně $\mathbb{D}$ vzhledem ke své funkční specifikaci ϕ, epistemické teorii $\mathcal{T}$ a etické teorii $\mathcal{E}$ právě když v každé situaci, ve které se může ocitnout, splňuje své funkční specifikace.*

Takto obecně pojatá umělá moudrost se tedy formálně skrývá v agentově funkční specifikaci Φ a je závislá na doméně $\mathbb{D}$ a teorii $\mathcal{T}$ a $\mathcal{E}$. Pokud jsou teorie $\mathcal{T}$ a $\mathcal{E}$ příliš jednoduché, nepostihující racionální chování agenta ve všech situacích a nezaručující vytvoření pragmatických či etických hodnot, může se agentovo chování jevit jako hloupé nebo neetické — ale to je v podstatě chyba návrhu. Agent totiž dělá přesně to, co mu jeho specifikace diktuje.

Tato definice dává smysl i z hlediska definice moudrosti tak, jak ji definovali starověcí filozofové a náboženští myslitelé. Tito zdůrazňovali maximální inteligenci, převyšující úroveň inteligence většiny lidí jako definiční vlastnost moudrosti s tím, že se jedná o zřídkavou kvalitu. Pokud odhlédneme od toho, že inteligence je opět slovo — kufr, naše definice vlastně ztotožňuje zřídkavou kvalitu inteligence se schopnosti naplňovat poslání agenta, definované v jeho funkčních specifikacích, ve všech situacích, se kterými se může setkat (a tedy, implicitně, dosahovat svých cílů etickým

Obr. 3. Théâtre D'opéra Spatial

způsobem). Co více chtít od agenta, jenž nemůže překročit svoji specifikaci? Všimněme si, že „zřídkavou kvalitu“ nejen umělé inteligenci dodává nikoliv pouze samotná schopnost dosahovat svých cílů, ale a způsob, jakým je jich dosahováno — etika.

Formálně definovaná umělá moudrost umožňuje hovořit o moudrosti extrémně jednoduchých kognitivních systémů, jakým jsou např. automatické otvírané dveře. Jsou moudré, protože otevřou dveře (vykonáním akce vytvoří pragmatickou hodnotu pro procházející osobu), kdykoliv rozeznají takovou potřebu (kognitivní schopnost), a chovají se eticky (pokud jsou zkonstruovány tak, že nikoho „nepřivřou“), a nic jiné se od nich nepožaduje. Složitější systém, jako třeba autonomní vozidlo, je také moudrý, protože (a pokud) pomocí kombinovaného efektu využití svých senzorů a motorů vytvoří pragmatickou a etickou hodnotu — doveze svého uživatele bezpečně k cíli.

Výhodou shora uvedené definice je skutečnost, že ukazuje, že moudrost není absolutní vlastnost, ale že závisí na schopnostech agenta, formálně popsaných v jeho „dvousložkové“ specifikaci, která zachycuje jak jeho jednání v každé situaci, tak i nutnost jednat tak, aby agent směřoval k naplnění svého poslání za dodržování etických hodnot. Nevýhodou je, že ji lze aplikovat pouze na formálně definované umělé systémy. Nám však nejde o to, poskytnout návod, jak takové systémy realizovat, ale jak je chápat a rozumět jejím možnostem a limitům.

5 Velké jazykové modely: Inteligence bez kognice

Jistou představu o tom, jak by mohly systémy generující moudrost v budoucnosti vypadat, nám dávají současné velké jazykové modely. Ukážeme, že tyto modely mají potenciál *správně používat své znalostí pomocí účelného chování — kombinovaný efekt kognice a jednání směřující k vytváření pragmatických a etických hodnot.*

Toto tvrdíme i přesto, že dle významného italo-britského filozofa Luciana Floridiho (2023) tyto systémy postrádají jakoukoliv inteligenci a porozumění, a nemají vůbec žádné kognitivní schopnosti. Tím pádem jsou velmi křehké (náchylné ke katastrofickým selháním), nespolehlivé (schopné dodat nesprávnou anebo vymyšlenou informaci), příležitostně schopné dělat elementární logické chyby v uvažování anebo v jednoduchých počtech. Floridi argumentuje, že v nich dochází k oddělení vazby mezi jednáním a inteligencí. Jde tedy o jednání bez inteligence, jak Floridi tyto systémy charakterizuje.

Mimochodem, je zajímavé si uvědomit, že v některých případech se mohou nedostatky velkých jazykových modelů, zmiňované Floridim, obracet ve výhody. To je případ systémů, generujících z textového popisu obrázky (jako je např. DALL-E (2023)). V případě obrázků totiž nelze jednoznačně určit, jestli je výsledný obrázek chybný, nebo padělek, anebo pouze měl systém (rozuměj: tvůrce obrázku) jinou představu než zadavatel. Obr. 3 (Théâtre d'Opéra Spatial 2023) ukazuje, že obrázek zadaný pouze textovým popisem může být oceněn jako kvalitní umělecké dílo. Shora uvedené nedostatky se v tomto kontextu mohou jevit jako jistý druh kreativity, vedoucí k nečekaným efektům, které nelze považovat za chyby, pouze za jakýsi „specifický umělecký vkus".

Vraťme se však zpět k velkým jazykovým systémům generujícím texty. Proč tedy tvrdíme, že tyto systémy, přes všechny shora uvedené nedostatky, mají potenciál vykazovat moudrost, alespoň v nějaké minimální míře? Uvedeme dva argumenty ve prospěch našeho tvrzení.

Prvním argumentem je samotný způsob práce těchto modelů,v jehož základem je *princip extrakce sémantiky ze syntaktických dat*. Tuto schopnost modelů Floridi (2023) sice zmínil, ale při svých úvahách nevzal v potaz povahu syntaktických dat, se kterými velké jazykové modely pracují. Tyto systémy jsou totiž schopné v překvapivé míře rozumět přirozenému jazyku, a pracovat s ním pomocí statistické analýzy různých vzorů, jež se nacházejí v kvantech syntaktických dat, na kterých je systém trénován. Porovnejme to s lidmi, kteří pracují s jazykem i jiným způsobem — sémantickým a kontextuálním usuzováním. Výsledek je však většinou tentýž — chápání významů zprostředkovanými jazykem (Agüera y Arcas, 2022). Již tento fakt samotný ukazuje, že tyto systémy nejsou zcela bez inteligence.

Náš druhý argument vychází z paradigmatu málo užívaného v oblasti umělé inteligence, kybernetiky anebo robotiky, ale o to známějšího v kognitivních vědách: *4E kognice* (viz např. (Newen, de Bruin & Gallagher, 2018) nebo (Extended mind thesis, 2022)). Toto paradigma postuluje, že kognice není pouhým vnitřním, individuálním procesem, nýbrž se jedná o emergentní proces, jež vzniká interakcí mezi mozkem, tělem, okolím a sociálním kontextem. Zkratka 4E naznačuje, že kognice je vtělená, vnořená, vykonávaná, a rozšířená (embodied, embedded, enacted, or extended) pomocí mimo-mozkových procesů a struktur.

Podívejme se, jestli a jak velké jazykové modely souzní s paradigmatem 4E.

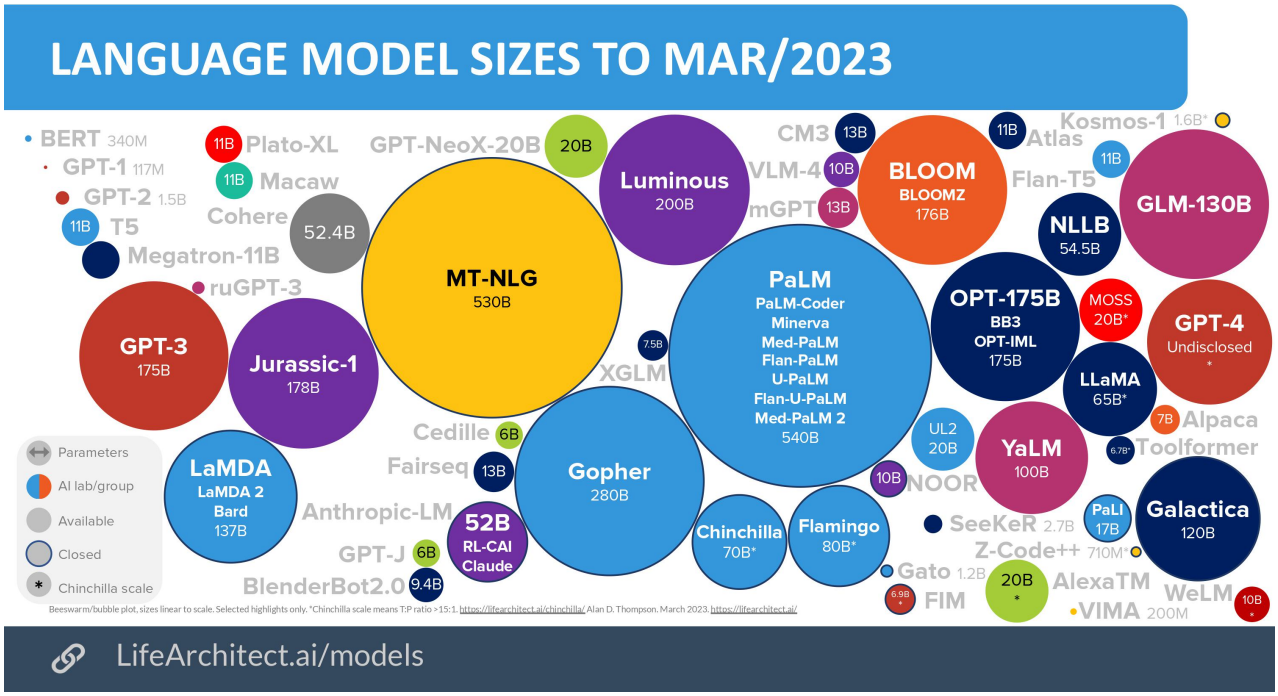
Vtělenost. Jazykové modely zřejmě nejsou fyzicky vtělené do reálného světa. Na druhé straně, učí se z obrovského množství dat a jejich porozumění jazyku vychází z kontextu, ve kterém se slova a fráze vyskytují. Tyto pocházejí z reálného světa, resp. přesněji, z představ lidí, kteří o tomto světe vytvářejí své texty. Takže lze říci, že prostřednictvím svých znalostí jsou tyto modely *nepřímo vtěleny* do reálného světa, protože poznávají svět pomocí interakce s texty pocházejícími s různých kontextů vyskytujících se ve skutečném světě.

Vnořenost. Za stejných důvodů jsou tyto modely *nepřímo vnořené,* zakotvené v reálném světě. Učí se asociovat slova s věty s kontextem, ve kterém se vyskytují a to jim dává význam v reálném světě.

Vykonatelnost. Způsob, kterým jazykové modely zpracovávají jazyk, lze chápat jako vykonatelný proces, protože tyto aktivně generují a manipulují své jazykové výstupy v závislosti na svých interních mechanismech. Tyto výstupy dále *nepřímo ovlivňují konání* lidí, kteří zadali vstupní data (prompt) do modelu.

Rozšířitelnost. Způsob, kterým velké jazykové modely reagují na svá zadání, zřejmě podstatně závisí na okolí systému, ze kterého „čerpá" své znalosti, a na promptu samotném, jež situuje systém v oblasti diskuse a poskytuje kontext pro porozumění jak vstupu, tak i výstupu. Tím je činnost sytému *nepřímo rozšiřovaná* a nasměrovaná, v závislosti na dotazu, do souvisejícího kontextu reálného světa.

Zastánci 4E kognice argumentují, že shora uvedené čtyři atributy mají vztah k inteligenci systémů, které splnění příslušných atributů vykazují. Je důležité si všimnout, že ve všech čtyřech případech atributů 4E kognice jsme hovořili o nepřímé vtělenosti, nepřímé vnořenosti, nepřímé vykonatelnosti, a nepřímé rozšířitelnosti. To jsou zcela odlišné atributy než ty, uvažované v klasickém případě 4E kognice. Nicméně, v tomto našem případě můžeme usuzovat z jednání systému přinejmenším o jakési *iluzorní inteligenci*, která nabízí iluzi inteligence vzhledem k danému prostředí, založenou na masivní agregaci dat z tohoto prostředí a výběru reakcí, jež závisejí na současném i minulém kontextu jednání systému. To je důvod, proč zcela nesouhlasíme s Floridiho závěrem, že ve velkých jazykových modelech je jednání odtržené od inteligence. Když, tak jde o jednání odtržené of reálné, bezprostřední kognice. (Mimochodem — tak jednají i lidé, pokud dospějí k rozhodnutí na základě „přemýšlení"). Iluzorní inteligence je v mnoha případech lepší než žádná inteligence. Otázkou zůstává, jestli součástí takové iluzorní inteligence je i nějaká forma vědomí. Pro úvahy, jestli velké jazykové modely mohou mít vědomí, viz práci (Chalmers 2022).



Obr. 4. Velikost jazykových modelů. Zdroj: https://s10251.pcdn.co/pdf/2023-Alan-D-Thompson-AI-Bubbles-Rev-7b.pdf

Znalosti využívané daným jazykovým modelem jsou ty znalosti o světě, které se model naučil analýzou textů získaných z Internetu. Všimněme si, že v takovém případě jazykové modely nevnímají svět „tak, jak vypadá“, tj. tak, jako ho vnímáme my lidé, ale pouze (neboli přesně) tak, jak se o něm, včetně o AI, píše. Text, který model vygeneruje, je požadovaná „umělá moudrost“, sémantika textu, který model vygeneruje, představuje pro nás ony *„pragmatické resp. etické hodnoty“*, a posléze samotný výpočetní akt konstrukce odpovědi odpovídá (resp. měl by odpovídat) „správnému používání znalostí modelu“. „Účelné chování“ je řízeno dotazem — model generuje text, který nejlépe souzní s odpovědí na daný dotaz. Reakce modelu současně nastavuje zrcadlo tomu člověku, který se ptá. Pokud se totiž nezeptá důkladně a nevysvětlí, co všechno chce vědět, včetně argumentů pro a proti, a nedožaduje se různých jiných, třeba i protichůdných názorů na danou věc, tak se dozví velmi málo, a povrchně, anebo dokonce dostane špatnou odpověď.

V reakci na iluzorní inteligenci velkých jazykových modelů můžeme tedy hovořit o *iluzorní moudrosti* takových systémů. Na rozdíl od kouzelníků, iluzionistů a jejich triků je zde situace příznivější. Pokud se nám něco na vygenerované moudrosti nezdá, anebo jen pro jistotu, můžeme požádat systém o vysvětlení, dodání třeba jiných, alternativních argumentů, a tak odhalit, jestli se jedná o iluzorní „moudrost“, anebo o moudrost, která je dobře zdůvodněná. To je současně návod, jak se stavět k používání současných velkých jazykových modelů.

V současné době se vyvíjí desítky různých jazykových modelů (viz. Obr. 4). Z hlediska technologie se tyto modely často velmi liší. Mohou používat různé architektury, trénovací postupy anebo techniky předzpracování dat. Z uživatelské hlediska jsou rozdíly mezi nimi jemnější. Vhodnost modelů se odvíjí od specifických úkolů, druhu a kvality trénovacích dat. Tato rozmanitost naznačuje, že velké jazykové modely a jejich architektury jsou robustní ve smyslu základní ideje, a naše výsledky k tomu dodávají, že tyto modely skutečně vládnou jakousi minimální, i když někdy iluzorní, inteligenci.

6 Proč rozvíjíme umělou inteligenci? Jaký účel má používání umělé inteligence?

Na základě všeho dříve řečeného začíná být zřetelná odpověď na dvě shora uvedené otázky. Umělou inteligenci rozvíjíme proto, abychom vyvinuli a poté mohli využívat nástroj na získávání (umělé) moudrosti — a to je současně i smysl jejího používání. Z definice DKIW hierarchie vyplývá, že moudrost, jako nejvyšší stupeň této hierarchie, nelze překonat. Někteří autoři se dokonce domnívají, že moudrost je více než inteligence (Jeste et al., 2020), ale to, zdá se, je otázkou definice inteligence — jestli je moudrost součástí inteligence

anebo ne.

Umělá moudrost se zdá být finální metou umělé inteligence. Jako celek to však není meta dosažitelná v konečném čase. Příklad matematiky, ve které prokazatelně existuje nekonečně, avšak spočetně mnoho teorií, to dokazuje. Velké jazykové systémy jsou prvními systémy AI, které naznačuji, ze v blízké budoucnosti budeme mit systémy, které budou generovat umělou moudrost a nebudou pouze pasivním „skladem" vědomostí (viz např. (Wiedermann & van Leeuwen, 2015b)). A to znamená zásadní posun společenského paradigmatu — od znalostní společnosti směrem k moudré společnosti. Umělá moudrost bude představovat společně s přirozenou moudrostí trvalý a smysluplný odkaz naším současníkům i potomkům a bude zvyšovat pravděpodobnost našeho a jejich přežití v zatím neznámých budoucích časech a místech.

7 Závěr

Umělá moudrost přestává být zajímavým slovním spojením, hodné filozofických úvah, ale začíná se jevit jako něco, co je na dosah dnešních technologií. Sam Altman, výzkumný ředitel firmy OpenAI, jež sestrojila a provozuje velký jazykový model GPT3 a GPT4, hovoří o tom, že finálním cílem OpenAI je všeobecná bezpečná umělá inteligence — AGI — s důrazem na slovo „bezpečná" (The Contradictions of Sam Altman, AI Crusader 2023). V kontextu moudrosti bychom slovo „bezpečná" nahradili slovem „etická". Proto nevidí důvod pro pozastavení vývojových prací na tomto projektu, jak požaduje Elon Musk a další vědci (Elon Musk, Other AI Experts Call for Pause in Technology's Development 2023). To je v souladu i s hlavní myšlenkou našeho příspěvku — směřovat k nacházení moudrosti pro naše současníky i potomky.

Poděkování: Tento příspěvek vznikl za částečné podpory TAČR v rámci projektu EBAVEL, registrační číslo CK04000150, a programu Strategie AV21 „Filozofie a umělá inteligence". Díky patří i systému ChatGPT za konzultace o práci velkých jazykových systémů.

Literatura

Agüera y Arcas, B. (2022). Do Large Language Models Understand Us?. Daedalus 2022; 151 (2): 183–197. doi: https://doi.org/10.1162/daed_a_01909

Chalmers, D. J. (2022). Could a Large Language Model be Conscious? https://philpapers.org/archive/CHACAL-3.pdf

DALL-E. (2023, March 29). In Wikipedia. https://en.wikipedia.org/wiki/DALL-E

DIKW pyramid. (2023, March 10). In Wikipedia. https://en.wikipedia.org/wiki/DIK_pyramid

Elon M. (2023). Other AI Experts Call for Pause in Technology's Development. (2023, March 29). The Wall Street Journal, https://www.wsj.com/articles/elon-musk-other-ai-bigwigs-call-for-pause-in-technologys-development-56327f?reflink=desktopwebshare_permalink

Extended mind thesis. (2022, December 1). In Wikipedia. https://en.wikipedia.org/wiki/ Extended_mind_thesis

Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models (February 14, 2023). Philosophy and Technology, 2023, Available at SSRN: https://ssrn.com/abstract=4358789

Jeste, D., Graham, S., Nguyen, T., Depp, C., Lee, E., & Kim, H. (2020). Beyond artificial intelligence: Exploring artificial wisdom. International Psychogeriatrics, 32(8), 993-1001. doi:10.1017/S1041610220000927

Large language model. (2023, March 29). In Wikipedia. https://en.wikipedia.org/wiki/Large_language_model

Minsky, M. (1998). Consciousness is a Big Suitcase: A talk with Marvin Minsky. https://www.edge.org/conversation/marvin_minsky-consciousness-is-a-big-suitcase

The Contradictions of Sam Altman, AI Crusader. (2023, March 31). The Wall Street Journal, htpps://www.wsj.com/articles/chatgpt-sam-altman-artificial-intelligence-openai-b0e1c8c9?reflink=desktopwebshare_permalink

Newen, A., De Bruin, L. and Gallagher, S. (eds). (2018). *The Oxford Handbook of 4E Cognition,* Oxford Library of Psychology (2018; online edn, Oxford Academic, 9 Oct. 2018)

Peano axioms. (2023, March 14). In Wikipedia. https://en.wikipedia.org/wiki/Peano_axioms

Théâtre d'Opéra Spatial. (2023, April 3). In Wikipedia. https://en.wikipedia.org/wiki/Théâtre_d'Opéra_Spatial

Vernon, D. (2021). Cognitive System. In: Ikeuchi, K. (eds) Computer Vision. Springer, Cham. https://doi.org/10.1007/978-3-030-63416-2_82

Wiedermann, J. (2017). Nový pohled na výpočty a umělá inteligence. In: Sborník přednášek konference *Kognice a umělý život 2017,* http://cogsci.fmph.uniba.sk/kuz2017/files/zbornik/Wiedermann.pdf

Wiedermann, J. a van Leeuwen, J. (2014). Computation as knowledge generation, with application to the observer-relativity problem. In: *Proc. 7th AISB Symposium on Computing and Philosophy: Is Computation Observer-Relative?*, AISB Convention 2014 (Goldsmiths, University of London), AISB, 2014

Wiedermann, J. a van Leeuwen, J. (2015a). What is Computation: An Epistemic Approach. (Invited talk). In: *Italiano, G. et al., (eds.). SOFSEM 2015: Theory and Practice of Computer Science.* LNCS 8939, Berlin: Springer, pp. 1-13

Wiedermann, J. a van Leeuwen, J. (2015b). Towards a Computational Theory of Epistemic Creativity. In: *Proc. 41st Annual Convention of AISB 2015.* London, pp. 235-242

Wiedermann, J. a van Leeuwen, J. (2017). Understanding and Controlling Artificial general Intelligenct Systems. In: *Proc. 10th AISB Symposium on Computing and Philosophy: Language, Cognition and Philosophy*, AISB Convention 2017, (University of Bath, UK), AISB

Wiedermann, J., & van Leeuwen, J. (2018). Epistemic computation and artificial intelligence. In Philosophy and Theory of Artificial Intelligence 2017 (pp. 215-224). Springer International Publishing.

Wisdom. (2023, March 16). In Wikipedia. https://en.wikipedia.org/wiki/Wisdom