Struktura, intence a automatismy řeči: davidsonovská teorie významu a velké jazykové modely

Petr Koťátko
Filosofický ústav AV ČR, v. v. i., Praha
kotatko@flu.cas.cz

Abstract:

Structure, Intentions and Speech Automatisms: Davidson's Theory of Meaning and Large Language Models

The article focuses on the possibilities and limits of AI systems, especially the large language models (LLMs), in confrontation with human linguistic competence. The primary topic is the ability of AI to perform various non-trivial communicative moves, including linguistic tricks and transitions between types of discourse differing in the principles of identifying utterance meanings. One of the key questions is how far one can proceed in describing these performances without applying intentional terms (while limiting oneself to purely "instrumental" language) and whether there is a fundamental difference in this respect between AI performances and human communication, particularly when its flow is carried by automatisms of speech. A significant part of the discussion takes place in confrontation with Davidson's project of a theory of meaning and his intentionalist "match-account" of utterance meanings.

Keywords: large language models, transitions between types of discourse, communicative intentions, speech automatisms, Davidson's theory of meaning

DOI: https://doi.org/10.46854/fc.2025.4r.589

0. Úvod

Předmětem následujících úvah budou možnosti a limity systémů umělé inteligence („artificial intelligence") typu velkých jazykových modelů („large language models")[1] v konfrontaci s lidskou jazykovou kompetencí a s jednou (svého času mimořádně vlivnou) teorií významu. Jistá část současných diskusí o AI se buď odehrává na půdě filosofie jazyka, nebo alespoň příležitostně

1 V dalším textu budeme pro stručnost (a pro přehlednost) zpravidla užívat zavedenou zkratku LLM v singuláru a LLMs v plurálu. Pro umělou inteligenci převezmeme obecně známou zkratku AI.

odkazuje k teoriím, argumentům či výzvám přicházejícím z této sféry. Průběžným vztažným bodem podstatné části našich úvah o možnostech LLMs bude Davidsonův projekt teorie významu na bázi Tarského teorie pravdy. Toto vztahování bude mít podobu jistých (poměrně volných) paralel a jistých (poměrně ostrých) konfrontací.

Předmětem polemické konfrontace bude především Davidsonův radikální intencionalismus, jeho reduktivní důsledky a způsoby, jakými nejen naše komunikativní akty, ale i možnosti LLMs překračují rámec nastavený touto redukcí. S využitím známého Davidsonova příkladu a jednoho protipříkladu se pokusíme demonstrovat schopnost systémů typu LLMs přecházet ve svých výkonech mezi typy diskursu s odlišnými principy konstituce významů promluv, jak k tomu dochází v lidské komunikaci.

Při popisu výkonů LLMs v těchto konfrontacích se (z metodologických, spíše než doktrinálních důvodů) omezíme na minimalistickou bázi: vyloučíme intencionální slovník[2] a nevstoupíme do diskusí o tom, zda jazykovým výstupům LLMs či chatbotů na nich založených lze přiznat status mluvních aktů typu „tvrzení, „otázka“ nebo „slib“. V závěrečné kapitole nicméně načrtneme jistou hypotézu o budoucím vyústění těchto debat.

1. Volba slovníku a zamýšlený postup

Nejprve bych rád zmínil dva z možných způsobů, jak popisovat aktuální nebo potenciální výkony systémů umělé inteligence, jako jsou velké jazykové modely.

(a) Prvním je *technicko-provozní* popis jednotlivých procedur, jako je tokenizace (tj. segmentace textu do krátkých úseků či prvků, jako jsou slova nebo morfémy), jejich kódování (převod segmentů – „tokenů“ na vektory v podobě n–tic reálných čísel) atd., nemluvě o popisu technické báze těchto výkonů, jako je konstrukce a způsoby fungování tzv. hlubokých neuronových sítí („deep neural networks“ – DNNs).

(b) Další možnost je popis výkonů LLMs pomocí *intencionálního* slovníku – jinými slovy způsobem, jakým jsme zvyklí mluvit o vlastních aktivitách. To znamená interpretovat výkony LLMs jako spojené s postoji typu „záměr vykonat tvrzení, že *p*“; „předpoklad (či přesvědčení), že jisté pokračování daného textu je vysoce pravděpodobné“; „snaha simulovat osobní styl vyjadřování pana *X*“ atd. To, samo o sobě, nás ještě nezavazuje k předpokladu,

2 Jsem si vědom toho, že v privátním i žurnalistickém diskursu je taková zdrženlivost běžně pokládána za nemístnou. Srov. též níže pozn. č. 29.

že LLMs jsou autentickými nositeli přesvědčení, záměrů a přání: může jít o metodické rozhodnutí zaujmout k LLMs a jejich výkonům takzvaný „intencionální postoj" („intentional stance") v Dennettově smyslu, to jest bez zmíněných implikací, prostě jen s nadějí, že to zvýší naši schopnost předjímat reakce systému na různé podněty.[3]

To jsou dvě možnosti, o nichž se zmiňuji jen proto, abych dodal, že se je nechystám využít (nanejvýš je příležitostně připomenu). V této stati půjdu jinou cestou:

(c) Výkony LLMs, o nichž budeme uvažovat, se pokusím popisovat pomocí *obecně instrumentálních* termínů, jako je „nastavení," „identifikace problému", „aktivace dostupné informační báze", „volba jednoho z pravděpodobných pokračování daného úseku textu" atd. Takový popis se, jak je zřejmé, neváže na aktuální technicko-provozní bázi těchto výkonů, a tedy na slovník typu (a), ani na intencionální slovník typu (b), jaký užíváme, když mluvíme o sobě.

Otázka bude znít, jak dalece si vystačíme s obecně instrumentálním slovníkem při popisu různých výkonů, u nichž budeme zvažovat, zda jsou (aktuálně či potenciálně) v dosahu možností systémů AI. Pokud jde o druh těchto výkonů, zaměříme se na různé komunikační tahy či strategie, včetně komunikačních triků. Zjistíme-li, že jejich popis pomocí slovníku typu (c) neobsahuje žádné mezery, k jejichž zaplnění jsou nutné výpůjčky ze slovníku typu (b), usoudíme z toho, že tyto výkony lze koherentně připsat systémům AI i v případě, že jim nepřiznáme intencionalitu (resp. zůstaneme v této věci neutrální).

Potom změníme perspektivu a zeptáme se, jak je to se stejnými či analogickými výkony ve sféře *lidské* komunikace: platí, že abychom je mohli připsat lidem, musíme je vyložit na intencionální bázi? Nebo z opačné strany: otázka zní, zda pro daný typ aktů, které jsme zvyklí připisovat lidem, najdeme možné a reálně uplatňované způsoby provedení, jejichž popis či výklad nejen nevyžaduje, ale vylučuje intencionální slovník (jeho nasazení by bylo

3 Daniel Dennett charakterizuje intencionální postoj a funkci, kterou může plnit, mimo jiné takto: "Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do." Srov. Dennett, D., *The Intentional Stance*. Cambridge, MA, The MIT Press 1996, s. 17.

zjevně zkreslující). A protože na tuto otázku budu (pokud se nezmění něco zásadního) vesměs odpovídat kladně, závěr bude znít, abychom to moc nepřeháněli v denunciačních poznámkách na adresu AI. Můžeme ji, po vzoru Emily Benderové, označit jako „stochastického papouška“,[4] a tak odkázat do patřičných mezí. Potom si ale pamatujme ten termín: mohl by se nám hodit, až budeme mluvit o sobě.

2. Diskuse o AI jako příležitost k úvahám o povaze jazykové komunikace a kompetence; první davidsonovská konfrontace

Samotné rozhodnutí pojmout diskusi o AI jako podnět k úvahám o nás není nijak výstřední. Řada současných autorů uvítala debaty o systémech AI jako příležitost, jak ukázat něco podstatného o povaze naší jazykové komunikace a naší jazykové kompetence,[5] případně o povaze jazyka samotného.[6] Nicméně dovolte, abych začal spíše historickou paralelou, která by, jak věřím, mohla být instruktivní. Mám na mysli Davidsonův projekt formální teorie významu na půdorysu Tarského teorie pravdy a jeho radikálně intencionalistické vyústění. Davidsonův intencionalismus v teorii významu a interpretace bude

4 Srov. Bender, E. M. – Gebru, T. – McMillan-Major, A. – Shmitchell, S., On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021.* Virtual Event Canada. New York, ACM 2021, s. 610–623. Dostupné na: https://doi.org/10.1145/3442188.3445922; [cit. 16. 9. 2025].

5 Neměli bychom pouštět ze zřetele dvojí možný výklad povahy jazykové kompetence. V úzkém smyslu jde o schopnost užívat a interpretovat věty jazyka nějakého společenství (tj. sociolektu) v souladu s jeho sémantickými konvencemi. V širším smyslu jde o obecnou schopnost úspěšně užívat jazykových prostředků, jako jsou nástroje singulární reference, predikace, kvantifikace atd., a interpretovat jejich užití nezávisle na tom, zda jde o užití a interpretaci konformní vůči jazykovým konvencím, a bez omezení na jeden jazyk jako zdroj těchto prostředků. Srov. níže pozn. č. 10. V kap. 5 a 8 budeme svědky toho, jak davidsonovské chápání funkce formální teorie významu à la Tarski postupuje od prvního pojetí k druhému.

6 Srov. např. "The exceptional performance of LLMs on sophisticated linguistic tasks, which were once considered strong indicators of human-like intelligence, raises intriguing questions regarding the nature of language processing, language understanding, and linguistic acts such as writing or speaking. These questions encompass broad and complex issues, and efforts to address them are still in their infancy." Mollo, D. C. – Millière, R., The Vector Grounding Problem. *Computer Science, Linguistics, Philosophy*, 4, April 2023. Dostupné na: https://arxiv.org/pdf/2304.01481; [cit. 16. 9. 2025]. Viz též Cappelen, H. – Dever, J., *Making AI Intelligible*. Oxford, Oxford University Press 2021; Piantadosi, S., Modern Language Models Refute Chomsky's Approach to Language. In: Gibson, E. – Poliak, M. (eds.), *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*. Berlin, Language Science Press 2024, s. 353–414; Grindrod, J., Large Language Models and Linguistic Intentionality. *Synthese*, 204, 2024, No. 71. Dostupné na: https://link.springer.com/article/10.1007/s11229-024-04723-8; [cit. 16. 9. 2025]. Millière, R., Language Models as Models of Language. In: Nefdt, R. – Dupre, G. – Stanton, K. (eds.), *The Oxford Handbook of the Philosophy of Linguistics*. Oxford, Oxford University Press (forthcoming). Dostupné na: https://www.researchgate.net/publication/383119605_Language_Models_as_Models_of_Language; [cit. 16. 9. 2025]; další odkazy v poznámkách níže.

přirozeným vztažným bodem podstatné části následujících úvah: půjde o to, do jaké míry jsou udržitelné univerzalistické nároky Davidsonova intencionalistického obrazu komunikace a co z toho plyne pro povahu, možnosti a meze výkonů LLMs, zvlášť pokud jim odmítneme připisovat intencionalitu.

Výchozí důvod ale spočívá v tom, že aspirace, které Davidson spojoval se svým projektem, odpovídaly motivaci, která dnes vede řadu filosofů k účasti v debatách o LLMs. Davidson, jak známo, nechápal svůj projekt jako cestu k vytvoření funkčního (v daném případě teoretického) aparátu, jehož chod bude generovat užitečné výstupy použitelné, dejme tomu, při překládání nebo při výuce jazyků. V našem kontextu je na místě dodat, že v takovém případě by nezáleželo na tom, zda operace vyžadované chodem tohoto aparátu (odvozování teorémů z axiomů) bude vykonávat člověk nebo systém AI. Důležité je, že Davidson sám svůj projekt chápal jako nepřímý způsob, jak zjistit něco podstatného o povaze jazykového významu, jazykové kompetence a způsobu fungování přirozených jazyků. Východiskem byl předpoklad, že plodným způsobem, jak zkoumat tyto věci, je položit si metateoretickou otázku, jaké podmínky musí splňovat formální teorie, má-li být schopna na konečné bázi (tj. z konečného počtu axiomů) generovat *něco tak silného* jako specifikace významů všech vět utvořitelných v nějakém přirozeném jazyce – a tak nám zpřístupnit to, co rodilým mluvčím poskytuje jejich jazyková kompetence.[7] To lze srovnat se současnými debatami o tom, jaké podmínky musí splňovat výkony systémů AI, abychom jim mohli přiznat status *něčeho tak silného*, jako jsou operace s jazykovými výrazy, které nesou význam, a to v rámci těchto operací samých, ne až sekundárně na základě toho, co jim (při výkladu výkonů AI) připisují jejich lidští interpreti.[8]

7 Srov. Davidson, D., Truth and Meaning a další stati obsažené v knize: Davidson, D., *Inquiries into Truth and Interpretation*. Oxford, Clarendon Press 1984. Na jedné straně nikdo, pokud vím, nezašel tak daleko, aby tvrdil, že davidsonovská formální teorie má sloužit jako věrná imitace naší přirozené jazykové kompetence, případně, že naše jazyková kompetence není nic jiného než internalizovaná teorie významu davidsonovského typu. Na druhé straně bylo možné předpokládat, že systematické odvozování významů vět z významů jejich složek a ze způsobů jejich spojení, jak k tomu dochází v davidsonovské teorii, zachycuje něco podstatného z povahy naší jazykové kompetence. Totiž její kreativitu, umožňující nám na konečné bázi smysluplně užít či interpretovat (principiálně) každou z nekonečného počtu vět utvořitelných v jazyce, jímž mluvíme.

8 Jinými slovy, otázka zní, za jakých podmínek můžeme jejich výkony uznat jako „bearing intrinsic meaning", a tedy jako projev něčeho, co se dá srovnat s jazykovou kompetencí. Srov. k tomu např. Mollo, D. C. – Millière, R., The Vector Grounding Problem (viz výše, pozn. č. 6); Havlík, V., Meaning and Understanding in Large Language Models. *Synthese*, 205, 2024, No. 1, s. 9. Dostupné na: https://doi.org/10.1007/s11229-024-04878-4; [cit. 16. 9. 2025].

3. Kontrast mezi založením a fungováním LLMs a tarskiovsko-davidsonovského formálního systému

Po této paralele je na místě poukázat na zásadní rozdíl mezi davidsonovskou formální teorií významu a způsobem, jakým systémy typu LLMs vykonávají své funkce. Základní složkou operační báze LLMs (a obecně všech „konekcionistických" systémů AI) není, jak známo, soubor axiomů jako zdroj možných inferencí ani aplikace symbolicky reprezentovaných jazykových pravidel, která by do nich byla vložena programátorem, ale (na rozdíl od „klasických" neboli „symbolických" systémů AI) zpracovávání rozsáhlého korpusu textů typicky (i když ne nutně) vytvořených lidmi, statistické vyhodnocování četnosti výskytů jednotlivých výrazů v jednotlivých kontextech a, na tomto základě, určování míry pravděpodobnosti možných pokračování textů, s nimiž je systém konfrontován.[9]

Na této bázi jsou pak LLMs schopny generovat pravděpodobná (a tedy věrohodně, resp. přirozeně působící) pokračování zadaných textů. To neznamená, že automaticky volí pokračování, které jim vychází jako nejpravděpodobnější: takové nastavení by s sebou neslo maximální míru předvídatelnosti, a tedy i uniformity (případně triviálnosti) jejich výstupů. Z tohoto důvodu se běžně volí nastavení („temperature"), které vede k nahodilému výběru jedné z vysoce pravděpodobných sekvencí navazujících vět. Při opakovaných zadáních téhož textu tedy můžeme očekávat různé varianty pokračování.

Rozvíjením této schopnosti v konfrontaci s rozsáhlým korpusem textů a v interakcích s lidskými protějšky dospěly LLMs, resp. chatboty na nich založené, k podivuhodným výkonům připomínajícím (ve svých vrcholech i propadech, o kterých ještě bude řeč) lidský komunikační provoz. Nabízí se otázka, jestli i plynulý tok *našich* řečových aktů, ať už v podobě textů, konverzací či monologů, není čas od času založen na mechanismech (nastavených tréninkem z dosavadní komunikační praxe), které nám, jako náš příspěvek do proudu psané či mluvené řeči, nabízejí věty, o nichž lze předpokládat, že budou přijaty jako přiměřené či přirozené pokračování toho, co předcházelo. A tato nabídka může mít stěží přirozenější základ než četnost

9 Jak upozornil jeden z anonymních recenzentů *Filosofického časopisu* (s odkazem na „poslední výzkumy firmy Anthropic"), současné LLMs jsou schopny nejen registrovat, ale i uplatňovat jazykové distinkce a vzorce, jejichž rozpoznání a aplikace přesahuje pouhou schopnost generovat pravděpodobná pokračování textů, a také překračuje běžné rozlišovací kapacity lidských uživatelů jazyka. Otázka zní, lze-li takové schopnosti vyvodit z „pouhého" vyhodnocování četnosti výskytů různých výrazů v různých kontextech, jak o něm byla řeč výše. Ať je to jakkoli, v následujících úvahách o možnostech AI v jistých komunikačních situacích zůstaneme u této klíčové operace: půjde nám o to, zda z této „minimální" báze je možné vyložit výkony (jisté netriviální komunikační tahy), na něž se zaměří náš zájem (viz níže kap. 6 a 9).

případů, v nichž se příslušné věty nebo jejich komponenty vyskytly a osvědčily v reakcích na příslušné podněty, jak jsme to stačili zaznamenat v reálném komunikačním provozu. V takových případech by zdroj našich výkonů a našich komunikačních dovedností neměl tak daleko k bázi, z níž čerpají výkony LLMs a chatbotů na nich založených. Na tomto místě to je předčasná úvaha – vrátíme se k ní v příhodnějších souvislostech.[10]

4. Vývoj Davidsonova projektu a způsob jeho zakotvení v reálné komunikaci; další kontrast a paralela s LLMs

Zpátky k naší konfrontaci s Davidsonovým programem: v této části bude jejím zdrojem jistý kritický bod vývoje projektu, Davidsonův způsob jeho překonání a jeho konsekvence pro úvahy o výkonech LLMs.

Máme zde tedy projekt teorie generující na konečné bázi teorémy pro každou z nekonečného počtu vět utvořitelných v jazyce *L*: teorémy formy Tarského T-vět

T: *„V“* je pravda tehdy a jen tehdy, když *p*,

kde na levé straně ekvivalence citujeme větu jazyka-objektu a na pravé straně specifikujeme její pravdivostní podmínky. Generování těchto teorémů má podobu jejich odvozování z axiomů vymezujících funkci jednotlivých složek vět a způsobů jejich spojení pomocí pojmů, jako je reference („John“ označuje Jana) nebo splňování (*x* splňuje „is a spy“ tehdy a jen tehdy, když *x* je špión).

Aspirace takového systému na roli teorie významu byly spojeny s jistými interními podmínkami, které měly, budou-li splněny, garantovat *interpretativnost* teorémů, tj. jejich schopnost specifikovat významy vět citovaných v jejich levé části. Tyto podmínky se ukázaly jako příliš slabé, jinými slovy

10 Otázka, zda je náš aktuální jazykový výkon motivován záměrem vyjádřit jistý propoziční obsah s jistou výpovědní silou nebo je pouhým produktem automatismu udržujícího v chodu proud řeči (ve smyslu našich úvah v kap. 7 a 8), může být za jistých okolností těžko zodpověditelná i pro samotné mluvčí. Za jiných okolností nedává smysl a automatismy zmíněného druhu nemají příležitost se spustit. Extrémní příklad: Čech a Němec se ocitnou ve společné cele v São Paulu: neumí portugalsky (až na pár slov), jeden neovládá jazyk druhého (až na pár slov), ale musí se dohodnout na společné akci. Nezbývá jim než čerpat z omezených zdrojů: užívat slova portugalštiny, o nichž se domnívají, že znají jejich význam a že stejný význam jim připíše partner; užívat slova vlastního jazyka, o nichž se domnívají, že jim porozumí partner; užívat slova partnerova jazyka, o nichž se domnívají, že již pochytili jejich význam. Z této báze pak vytvářejí své věty, s tím, že prázdná místa vyplňují mimikou a gesty. Půjde o věty v plném smyslu: věty, jejichž složky plní funkce singulárních termínů, predikátů, kvantifikátorů, pravdivostně funkčních spojek, indikátorů výpovědní síly („illocutionary force“) atd. Srov. výše pozn. č. 5.

nestačily k vyloučení zjevně neinterpretativních teorémů.[11] Následné diskuse nevedly k nalezení nějaké další interní restrikce, která by konečně překlenula distanci mezi materiálními ekvivalencemi typu T-vět a plnohodnotným vymezením větného významu.

Výsledkem bylo přiznání, že formální struktura typu davidsonovské teorie, jakkoli vztažená k reálnému světu prostřednictvím pojmů jako reference, splňování nebo pravda, nezaručuje, ani při splnění zmíněných podmínek, generování významových přiřazení (tj. interpretativních teorémů) a nemůže tomu, kdo s ní operuje, ať už by šlo o člověka nebo o klasický („symbolický") systém AI, sloužit jako substitut jazykové kompetence. K tomu musí být zasazena do fundamentálnějšího rámce. Pokud jde o hierarchii teorií, znamená to zahrnout formální tarskiovsko-davidsonovskou teorii do neformální teorie interpretace jako její podřízenou součást. Pokud jde o budování teorie, utváření a ověřování jejího sémantického potenciálu, musí se odehrávat v reálném komunikačním provozu, v interpretační práci v terénu zaměřené na promluvy reálných mluvčích.

Zde se nabízí přímočará, i když jen velmi volná paralela s LLMs. Jak jsme si už připomněli, klíčovou součástí jejich formování je učení na textech reálných uživatelů jazyka a účast v reálných komunikačních výměnách. Právě tak formování, ověřování a fungování davidsonovské teorie má být ponořeno do reálného chodu komunikace: má postupovat „zdola nahoru" (tj. být „data-driven"), jak jsme si zvykli říkat o LLMs.

Přejdeme-li však ke způsobu, jakým se to děje, místo paralely nastupuje kontrast. Určujícím rysem této fáze vývoje Davidsonova programu je předpoklad principiální korelace mezi sémantickými charakteristikami vět a psychologickými postoji jejich uživatelů v podobě nového, tentokrát neformálního omezení uvaleného na davidsonovskou teorii, známého jako „restrikce s ohledem na propoziční postoje" („propositional attitude constraint"). A obecněji, určující roli zde přejímá Davidsonův radikální intencionalismus, nad nímž si položíme dvě otázky:

11 Původní vymezení podmínek interpretativnosti teorémů lze shrnout takto: Teorém pro větu *V* jazyka *L* specifikuje význam *V* v *L* tehdy a jen tehdy, když je pravdivý a odvoditelný v rámci teorie, v níž jsou odvoditelné pravdivé teorémy pro všechny věty *L*. Zvažme ale protipříklad: T_n: „John is a spy" je pravda tehdy a jen tehdy, když Jan je špión a Jirásek napsal *Temno* nebo Jirásek nenapsal *Temno*. Jak konstatovali účastníci diskusí na toto téma, neplatí, že zjevná *neinterpretativnost* pravdivých teorémů typu T_n by se musela projevit v *nepravdivosti* jiných teorémů odvoditelných v teorii, v níž by byl odvoditelný T_n. Teorém T_n tedy může splňovat podmínky interpretativnosti zmíněné výše, včetně holistické restrikce. Závěr zní, že podmínky jsou málo restriktivní. Viz k tomu např. Evans, G. – McDowell, J. (eds.), *Truth and Meaning: Essays in Semantics*. Oxford, Oxford University Press 1976 (zejména stati J. Fostera a B. Loara). Stručný výklad davidsonovské teorie a diskusí, které vyvolala, viz např. v: Koťátko, P., *Interpretace a subjektivita*. Praha, Filosofia 2006, s. 413–418.

(a) Které z jeho parametrů se mohou promítnout do výkonů systémů AI i v případě, že jim samým nepřiznáme intencionalitu?

(b) Hraje opravdu intencionalita tak všudypřítomnou a neeliminovatelnou roli v lidské komunikaci? Nebo je nezanedbatelná část lidských komunikativních výkonů zcela srovnatelná s výkony systémů AI, popisovanými v ne-intencionálním slovníku?

5. Davidsonův radikální intencionalismus

Vraťme se tedy k vývoji Davidsonova programu. Přechodem na širší (a fundamentálnější) bázi, do sféry teorie interpretace, máme získat skutečně funkční kritérium schopnosti formální teorie Tarského typu plnit funkci teorie významu, tj. generovat interpretativní teorémy. Debaty na toto téma se ubíraly dvěma směry: v obou hrál ale klíčovou roli intencionalistický slovník. Jedna linie směřovala k vymezení podmínek schopnosti teorie zachytit jazyk užívaný nějakým společenstvím, například za pomoci pojmu mínění něčeho promluvou („speaker's meaning", resp. „utterer's meaning"), převzatého z Griceovy intencionální sémantiky v kombinaci s Lewisovým intencionálně definovaným pojmem konvence.[12] Druhá linie vztahovala teorii výlučně k samotnému komunikačnímu provozu (k interakcím mezi mluvčími a adresáty) a předpoklad konvenčně zakotveného sociolektu v ní nehrál zásadní roli. Zaměříme se na tento proud, v němž se stále výrazněji (a radikálněji) angažoval sám Davidson.

Základní teze zněla, že T-věty fungující jako hypotézy o významech vět užitých v reálné komunikaci jsou interpretativní právě tehdy, když do sebe zapadají s našimi psychologickými hypotézami o záměrech a přesvědčeních mluvčích, a to způsobem, který nám umožňuje porozumět jazykovému a na ně navazujícímu mimojazykovému jednání mluvčích.[13] Jinými slovy,

12 Obecné schéma: Jazyk *L* reprezentovaný davidsonovskou teorií je skutečným jazykem („actual language") společenství *S* tehdy a jen tehdy, když pro každou větu *V* jazyka *L* a specifikaci pravdivostních podmínek *p* platí, že teorie generuje teorém formy „*V*" je pravda tehdy a jen tehdy, když *p*," tehdy a jen tehdy, když v *S* platí konvence (v Lewisově smyslu) užívat *V*, když mluvčí míní (v Griceově smyslu), že *p*, případně konvence interpretovat užití *V* jako *prima facie* evidenci, že mluvčí míní, že *p*. K Lewisově definici konvence srov. zvl. Lewis, D., Languages and Language. In: Lewis, D., *Philosophical Papers*. Vol. 1. Oxford, Oxford University Press 1983. Českou verzi s komentářem viz v: Koťátko, P., *Interpretace a subjektivit*a, c.d., s. 438–442. Ke Griceově intencionální sémantice srov. tamtéž, s. 430–433.

13 Spojíme-li tuto formu „propositional attitude constraint" s restrikcemi zmíněnými v pozn. č. 11, dostaneme tezi, že interpretativni hodnotu teorému davidsonovsko-tarskiovské formální teorie pro jazyk *L* může garantovat jen soubor podmínek typu: (a) *T* je pravdivý; (b) *T* je odvoditelný v rámci teorie generující pravdivé teorémy pro všechny věty *L*; (c) *T* přispívá, ve spojení

když přispívají k tomu, aby nám toto komplexní jednání dávalo smysl jako koherentní jednání racionálních bytostí. To, obecně vzato, dobře ladí s Wittgensteinovým pojetím významu, vyjádřeným mimo jiné v tezi: „To, co říkáme, nabývá svůj význam ze zbytku našeho jednání."[14] Jinými slovy: význam našich promluv je dán jejich funkcí v rámci (jazykového i mimojazykového) jednání, do něhož jsou zasazeny. U Davidsona k tomu podstatně patří, že to, co se na této bázi konstituuje jako význam promluvy, nevyžaduje potvrzení (ratifikaci) jazykovými konvencemi. Řečeno jinak: máme-li při interpretaci promluvy (a zvažování naší reakce na ni) volit mezi významem, v jehož světle nám bude komplexní (jazykové a mimojazykové) chování mluvčího dávat smysl jako koherentní jednání racionální bytosti, a významem, který dostaneme aplikací jazykových konvencí na promluvu, máme se bez váhání rozhodnout pro první verzi (více k tomu níže v kap. 7).

6. Případ paní Malapropové jako test pro možnosti LLMs

Zvažme známý příklad takové situace a položme si otázku, jak by se s ní mohl vypořádat systém typu LLM, s jeho operačními možnostmi, a přitom způsobem, který by (ve svém výsledku) odpovídal Davidsonovu postulátu. Půjde o slavný případ paní Malapropové, který si Davidson vypůjčil ze hry Richarda Brinsleyho Sheridana *The Rivals* (1775). Představme si salonní společnost debatující nad textem nějaké básně. Jedna z přítomných, paní Malapropová, hledá příležitost, jak se vmísit do rozpravy, přestože není právě doma v literárněvědné terminologii. S tímto záměrem ukáže na nějaký verš a řekne: „This is a nice derangement of epitaphs." Z kontextu je všem přítomným zřejmý zamýšlený obsah: nechtěla tvrdit, že na daném místě nachází pěknou destrukci náhrobních nápisů, ale upozornit na pěkné uspořádání básnických přívlastků.

Jedna z možných reakcí ze strany účastníků konverzace by zněla takto: „Sdílím vaši zálibu v náhrobních nápisech, ale žádné tu nenacházím." Pokud

(*Pokrač. pozn. č. 13*) s našimi hypotézami o propozičních postojích uživatelů *L* a o okolnostech je jich promluv, k porozumění komplexnímu (jazykovému i mimojazykovému) chování uživatelů *L*. Psychologickým hypotézám (připisování postojů, jako je přesvědčení nebo záměr) přitom u Davidsona náleží metodologická priorita v tom smyslu, že vyjde-li nám jazykové a s ním spojené mimojazykové jednání interpretované osoby jako iracionální (jako jednání nevykazující žádný racionální vzorec), máme přehodnotit své sémantické hypotézy, abychom udrželi psychologické hypotézy, dokud je nevyvrátí přesvědčivá evidence. Srov. např. "This method is intended to solve the problem of the interdependence of belief and meaning by holding belief constant as far as possible while solving for meaning." Davidson, D., Radical Interpretation. In: týž, *Inquiries into Truth and Interpretation*, c.d., s. 137.

14 "Our talk gets its meaning from the rest of our proceedings." Wittgenstein, L., *On Certainty*. Oxford, Blackwell 1969, § 229/30e.

ji zvolíte, podtrhnete trapnou záměnu, jíž se dopustila paní M., sám se projevíte jako neochvějný ctitel jazykových konvencí, a tedy, v očích části přítomných, jako odpovědný, zásadový a zároveň duchaplný člověk, v očích jiných jako hulvát, v každém případě ale ne jako vstřícný interpret se zájmem o dorozumění. Vstřícná (a komunikativně funkční) reakce by například zněla: „Který z těch přívlastků je podle vás zvlášť působivý?"

Obecně vzato, situace nás staví před volbu mezi respektem (loajalitou, submisivitou) vůči *jazykovým konvencím* a respektem (vstřícností, empatií) vůči *člověku*, našemu partneru v komunikaci. Otázka zní: v jaké podobě může takové či analogické dilema vyvstat před systémem typu LLM? Nejspíš budeme mít silné zábrany mluvit zde o volbě mezi loajalitou a empatií. Ale dilema je možné vymezit čistě instrumentálně:[15]

(a) Jedna možnost je zvolit reakci v podobě věty, která patří mezi pravděpodobná pokračování tahu, jímž byla promluva paní Malapropové, tj. její vyslovení věty „This is a nice derangement of epitaphs".

(b) Druhou možností je registrace faktu, že pravděpodobnost výskytu slov „derangement" a „epitaphs" v dané fázi konverzace[16] se blíží nule. Dalším krokem je určení okruhu slov, jejichž výskyt na daném místě (v rámci užité věty a v dané fázi konverzace) vychází jako vysoce pravděpodobný. Následujícím krokem je výběr dvojice z této nabídky, splňující další podmínku: má jít o slova nejpodobnější těm, která byla aktuálně užita. „Arrangement" a „epithets" budou bezpochyby slibní kandidáti, nejspíš zdaleka nejvhodnější ze všech. Z jejich dosazení do původní věty (na místa uvolněná slovy, která neobstála v pravděpodobnostním testu) vzejde věta „This is a nice arrangement of epithets" a nahradí původní verzi v pozici věty, s níž se pak poměřuje pravděpodobnost možných pokračování: stane se tedy vztažným bodem pro volbu následující věty.

Volba mezi (a) a (b) může záviset na tom, co je v daném systému aktuálně (v dané fázi „učení") nastaveno jako „normální postup", buď obecně, nebo pro daný typ situace. Případně může záviset na odhadu konsekvencí každého ze zvažovaných postupů, konkrétně na tom, jak bude ta či ona odpověď na promluvu pravděpodobně přijata ostatními účastníky konverzace (odhad bude vycházet z analogických situací zaznamenaných v minulosti). Kritériem volby pak může být to, nakolik toto přijetí ostatními posílí či oslabí pozici mluvčího (v našem případě systému AI), a tedy rozšíří či zúží jeho možnosti v dalším vývoji komunikace (a vzájemných interakcí vůbec).

V případě AI zřejmě nebudeme předpokládat, že preference pro jeden z těchto předvídatelných důsledků bude zakotvena emocionálně (tj. touhou

15 Zde: s ohledem na standardní operační bázi výkonů LLMs (viz výše kap. 3).
16 Tj. po předchozích větách typu "What would you say about this poem?"

po společenském úspěchu nebo obavou z neúspěchu), případně morální volbou (mezi loajalitou *k pravidlům* a vstřícností či empatií *k člověku*). V našem případě si vystačíme s předpokladem, že chod systému je nastaven na dosažení kladné, potvrzující reakce (na tom je ostatně založeno tzv. posilující učení z lidské zpětné vazby, známé pod zkratkou RLHF[17]).

I když tedy odmítneme pohlížet na systém jako na subjekt propozičních postojů nebo nositele emocí a morálních intuic, které jsme vyhradili sobě, nestává se tím nutně indiferentním vůči volbám a inertním vůči podnětům, které vycházejí z dané konverzační situace a navozují nutnost rozhodnutí, jako je volba v kauze paní Malapropové.[18]

7. Konfrontace automatismů lidské komunikace a fungování LLMs: „nastavení" na pravděpodobné přijetí našich promluv

Ostatně vyzkoušejme pohled z druhé strany. Jde o to, jak je to v těchto věcech *s námi* – například zda v typické rutinní konverzaci nesledujeme známé, v předchozím tréninku vyzkoušené, konverzační trasy směřující ke statusu (či ocenění) typu „vstřícný" nebo „korektní" nebo „precizní" nebo „vtipný" podle (rutinního) odhadu pravděpodobného přijetí našich promluv v daném prostředí v dané fázi konverzace. Další otázka by zněla, zda tento odhad nemá základ v osobní statistice četnosti užití těch a těch obratů v těch a těch kontextech a četnosti těch a těch reakcí na tyto výkony. Mělo by být zřejmé, že osobní statistikou zde nemyslím výsledek metodicky prováděných operací, ale produkt zkušeností, které se ukládají a kumulují v naší paměti.

Otázka zní, zda tento popis naší role v každodenním konverzačním provozu je pustá karikatura, nebo jen věcně (a tedy bez iluzí) zachycuje *jednu z možných cest*, jakými z nás vycházejí takzvané rozumné, vstřícné, případně vtipné repliky v běžných konverzacích. Ale dokonce i tyto aspirace, které nelze zaměňovat s komunikativními intencemi v úzkém smyslu (vyjádřit

17 „Reinforcement learning from human feedback" je jedním z druhů „jemného dolaďování" („fine tuning"), sloužících k přizpůsobení funkcí systému našim preferencím. Instruktivní výklad nabízí např. výše citovaná stať: Mollo, D. C. – Millière, R., The Vector Grounding Problem, c.d., s. 23.

18 Jeden z anonymních recenzentů *Filosofického časopisu* poukázal na to, že v takových situacích se „nejnovější model americké společnosti OpenAI (GPT4.5) přiklání k variantě respektovat mluvčí; čínský model DeepSeek R1 naopak k loajalitě vůči konvencím…". To mi připomnělo dávnou epizodu: jeden z oponentů mé někdejší kritiky Davidsonovy teorie významu z pozice (tehdy už dosti vlažného) normativismu mi můj respekt ke konvencím vyložil jako pozůstatek z časů, které jsem prožil v totalitě. Davidson sám poznamenal, že „Koťátko, stejně jako [R. M.] Hare, věří, že existuje nějaký druh zákona nebo pravidla, podle nějž platí, že kdykoli jsou určitá slova vyslovena určitým způsobem, je bez ohledu na záměr [mluvčího] vykonáno tvrzení." Davidson, D., Odpowiedz Petrowi Kotatce. In: Zeglen, U. (ed.), *Dyskusje z Donaldem Davidsonem o prawdzie, jezyku i umysle*. Lublin, KUL 1996, s. 341.

jistý propoziční obsah s jistou výpovědní silou) se mohou rozpustit v proudu řeči, v němž jedna věta funguje jako spouštěč pro druhou jako své přirozené, samo se nabízející pokračování, a plynulý chod tohoto mechanismu nevyžaduje žádnou intervenci komunikačních záměrů v úzkém nebo širším smyslu: jde jenom o to, aby řeč nestála.[19] Už jenom připustit, že jsem právě popsal jednu z možných podob našeho komunikačního provozu, může působit jako projev krajního cynismu. Chtěl bych jen připomenout, že spisovatelé a dramatici vybavení (či postižení) zvláštní vnímavostí ke způsobům, jakými funguje náš jazyk, občas zobrazují naše každodenní konverzace právě v tomto duchu.[20]

Shrňme si pro předběžnou bilanci část cesty, kterou jsme zatím prošli. Na začátku bylo naše (programové) rozhodnutí obejít se při popisu jazykových výkonů LLMs bez intencionálních termínů typu „záměr" nebo „přesvědčení", případně „vstřícnost" nebo „vzájemnost", které jsme si vyhradili pro sebe. V případě AI jsme proto zvolili čistě instrumentální popis, zaměřený na zachycení chodu systému v termínech jako „nastavení", „registrace", „statistika četnosti výskytů různých výrazů v různých kontextech", „odhad pravděpodobného přijetí jednotlivých výstupů" atd. Výsledkem bylo, že jsme, proti původnímu plánu, náhle stáli před problémem, jak vyvrátit podezření, že jsme právě našli adekvátní popis *našich vlastních* výkonů v jistých komunikačních situacích, jejich zdrojů a jejich směřování – a situace se obrátila.

19 Nemám zde na mysli poruchy řeči a jejich projevy, jako je tzv. hyperfluence – patologicky naddimenzovaná míra plynulosti či spádu toku řeči (nad 200 slov za minutu – oproti běžné fluenci ve výši 120 slov za minutu), jejíž komunikační přínos bývá (klinicky i laicky) hodnocen jako „bezobsažný" či „bezúčelný". Mám na mysli projevy běžně hodnocené jako doklad mimořádné výmluvnosti, suverénní orientace v dané sféře (případně v několika těžko vymezitelných sférách současně), hlubokého vhledu do problému atd., s mírou fluence, která vyvolává obdiv. K hyperfluenci a dalším poruchám fluence jako projevům afázie viz např. Weiner, M. – Neubecker, K. – Bret, M. – Hynan, L., Language in Alzheimer's Disease. *The Journal of Clinical Psychiatry*, 69, 2008, No. 8. Valerie M. Balesterová nabídla sociolingvistické (spíše než psychiatrické) vymezení pojmu hyperfluence jako formy (mluveného i psaného) projevu vyznačující se silně naddimenzovaným a často věcně neobhajitelným výskytem expertních termínů či frází a přistoupila k ní jako k přechodné, a ne nutně neproduktivní fázi osvojování jazyka daného společenství, prostředí či oboru. Viz Balester, V. M., Hyperfluency and the Growth of Linguistic Resources. *Language and Education*, 5, 1991, No. 2, s. 81–94.

20 V knize *O čem se vypráví* (Praha, Filosofia 2023) uvádím příklady tohoto i opačného druhu z textů Franze Kafky (monology advokáta v *Procesu*, nebo starosty v *Zámku*) a Samuela Becketta (dialogy z románové *Trilogie* nebo z pozdních povídek a novel); výmluvným příkladem jsou i nekonečné monology v hrách a prózách Thomase Bernharda, nebo předvádění vzorců „domácí konverzace" v Ionescově *Plešaté zpěvačce*. Srov. též úvahy Beckettova vypravěče (v *Nepojmenovatelném*; viz k tomu: Koťátko, P., *O čem se vypráví*, c.d., s. 275–277) o tom, kdo právě mluví jeho ústy a kdo je skutečným subjektem myšlenek a prožitků, které procházejí jeho myslí – a jeho neschopnost rozhodnout v této věci. Je na nás, budeme-li takové texty vnímat jako literární výstřednosti vedené snahou upoutat náš zájem – nebo si je vyložíme tak, že lidé mimořádné vnímavosti a zároveň mimořádné autorské poctivosti a důslednosti nám tu předvádějí (ukazují, spíše než vysvětlují), jak na tom naše řeč (také někdy) je – a my s ní.

Namísto otázky, co musí splňovat systémy AI, aby se ve svých výkonech *přiblížily* lidským mluvčím a interpretům, nastoupila otázka, jak vymezit rysy, jimiž se od AI v těchto rolích *zásadně lišíme* – v tom smyslu, že jsou prokazatelně přítomné ve všech případech, které lze označit jako *mezilidskou* jazykovou komunikaci, zatímco chybí ve výkonech AI. Závěr zněl, že komunikativní intence (v užším ani širším smyslu) nejsou slibným kandidátem na tuto funkci.

8. Reduktivní důsledky Davidsonova pojetí významu promluvy a případ komunikačního tahu s „rozbíhajícími se intencemi"

Po této diskusi, inspirované slavným Davidsonovým příkladem, se soustřeďme na koncept významu promluvy („utterance meaning"), který měl tento příklad demonstrovat. Jak jsme už měli příležitost zaznamenat, význam promluvy v davidsonovském pojetí není dán aplikací jazykových konvencí na promluvu a její kontext. Jeho vlastním zdrojem je shoda komunikativní intence a interpretace: promluva je tvrzením, že *p*, právě tehdy, když tak byla míněna mluvčím a je tak interpretována adresátem.[21] Význam promluvy, konstituovaný touto shodou, nevyžaduje ani dodatečně žádnou ratifikaci ze strany jazykových konvencí – je-li vlastním cílem komunikace dorozumění, spíše než potvrzení naší loajality ke konvencím, případně udržování jazykového aparátu v chodu.[22]

V následující debatě budeme mít příležitost konstatovat, že Davidsonův radikálně intencionalistický přístup k významu promluvy je, kvůli svým uni-

21 Srov. např. "Meaning, in the special sense in which we are interested when we talk of what an utterance literally means, gets its life from those situations in which someone intends (or assumes or expects) that his words will be understood in a certain way, and they are. In such cases we say without hesitation: how he intended to be understood, and was understood, is what he, and his words, literally meant on that occasion. There are many other interpretations we give to the notion of (literal, verbal) meaning, but the rest are parasitic upon this." Davidson, D., The Social Aspect of Language. In: McGuiness, B. – Oliveri, G. (eds.), *The Philosophy of Michael Dummett*. Dordrecht – Boston – London, Kluwer Academic Publishers 1994, s. 11–12. Nebo: "Meaning of whatever sort rests ultimately on intention." Davidson, D., Locating Literary Language. In: Dasenbrock, R. W. (ed.), *Literary Theory after Davidson*. University Park, The Pensylvania State University Press 1993, s. 298; nebo: "In the end, the sole source of linguistic meaning is the intentional production of tokens of sentences. If such acts did not have meaning nothing would." (tamtéž).

22 Srov. např. "Beliefs, desires and intentions are a condition of language…, convention is not a condition of language." Davidson, D., Communication and Convention. In: týž, *Inquiries into Truth and Interpretation*, c.d., s. 280. Vyústěním této linie úvah o komunikaci je závěr: "I conclude that there is no such thing as a language, not if language is anything like what many philosophers and linguists have supposed. There is therefore no such thing to be learned, mastered or born with." Davidson, D., A Nice Derrangement of Epitaphs. In: Lepore, E. (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Cambridge, Blackwell 1986, s. 446.

verzálním aspiracím, silně *reduktivní*, a to i s ohledem na možné intencionální báze našich komunikačních strategií. Zvážíme jistý protipříklad (v podobě komunikačního triku) a bude nás zajímat (v kap. 9), jak se s případy tohoto druhu mohou vypořádat systémy AI, když jsou s nimi konfrontovány v komunikaci, případně na jaké bázi je mohou uplatňovat jako vlastní komunikační tahy.

Nastal tedy čas poukázat na davidsonovské přecenění role komunikativních intencí (jejichž obecnou bází je záměr být interpretován tak a tak) v reálné komunikaci – to bude důležité pro naše úvahy o možnostech dostupných systémům AI i v případě, že jim odmítneme přiznat intencionalitu. V konfrontaci s davidsonovským radikálním intencionalismem bych chtěl poukázat na dvě věci:

(a) V předchozí části už byla řeč o situacích, v nichž tím, co žene mluvčího k dalším a dalším výkonům v řetězu promluv, nejsou komunikativní intence v přísném smyslu, tj. záměry vyjádřit jisté propozice s jistou ilokuční silou, ale prosté nutkání udržovat řeč v chodu, pouhá setrvačnost, neschopnost vystoupit z toku řeči. Případně snaha vmísit se do rozběhlé konverzace, i když do ní nemám čím přispět a neovládám ani její slovník. Představme si, že promluva paní Malapropové z Davidsonova příkladu (viz kap. 6) není vyvolána konkrétní komunikativní intencí, kterou jsme jí připsali (záměrem vykonat jisté tvrzení o básnických přívlastcích), ale prostě jen nutkáním nezůstat stranou salonní zábavy, vplynout do ní i do jejího stylu užitím znalecky znějících slov – tak jako Molièrův doktor (ve *Zdravém nemocném*) vrší bez ohledu na významy (či jejich absenci) expertně znějící výrazy, ne proto, aby vykonal tvrzení s jistým propozičním obsahem, ale aby ohromil pacienta svou učeností a vytáhl z něj, co se dá. To jsou případy, kdy komunikativní intence mluvčího v přísném smyslu, s nímž pracuje Davidson, vůbec nevstupuje do hry.[23]

23 Zároveň jde o příklady fenoménu běžně označovaného jako „bullshit" – o promluvy vykonávané bez zřetele k rozdílu mezi pravdou a nepravdou. *Locus classicus* je proslulá kniha Harryho G. Frankfurta: *On Bullshit*. Princeton, Princeton University Press 2005. Někteří autoři aplikovali tento termín na výkony LLMs a chatbotů na nich založených; srov. Hicks, M. T. – Humphries, J. – Slater, J., ChatGPT is Bullshit. *Ethics and Information Technology*, 26, 2024, article number 38, published online 8. 6. 2024. Dostupné na: https://doi.org/10.1007/s10676-024-09775-5; [cit. 16. 9. 2025]. Běžnější označení pro výkony LLMs indiferentní k požadavku pravdivosti je „hallucination". K rizikům s tím spojeným viz např. Coeckelbergh, M., LLMs, Truth and Democracy: An Overview of Risks. *Science and Engineering Ethics*, 31, 2025, article number 4. Dostupné na: https://link.springer.com/article/10.1007/s11948-025-00529-0; [cit. 16. 9. 2025].

(b) Přejděme k případům, kdy, na rozdíl od předchozího, mluvčí má zcela určitou komunikativní intenci (tj. záměr vykonat mluvní akt s jistým propozičním obsahem a výpovědní silou). V takových případech může důsledná aplikace Davidsonova vymezení významu promluvy a komunikativního úspěchu vést, paradoxně, k simplifikaci intencionální stránky komunikace, k redukci možných podob a kombinací záměrů spojených s promluvou. Rozveďme nejprve intencionální bázi komunikativního aktu a jeho recepce: přijmeme-li davidsonovské pojetí významu promluvy jako shody komunikativní intence a interpretace, můžeme rozlišit několik parametrů komunikativního úspěchu v podobě naplnění intencí mluvčího a adresáta. Tyto intence koincidují v tom smyslu, že naplnění jedné vyžaduje naplnění ostatních (vyžaduje komunikativní úspěch, jenž zahrnuje naplnění všech intencí z našeho výčtu):

I_1: záměr mluvčího vykonat konkrétní mluvní akt *A* (identifikovaný propozičním obsahem a ilokuční silou)
I_2: záměr mluvčího, aby jeho promluva byla interpretována jako vykonání aktu *A*
I_3: záměr adresáta identifikovat akt vykonaný v dané promluvě mluvčím
I_4: záměr adresáta identifikovat záměr I_1
I_5: záměr adresáta identifikovat záměr I_2

Soustřeďme se na vztah mezi I_1 a I_2 a představme si, že Pavel říká Karlovi „Martialis psal skvělé epigramy" se záměrem vykonat tvrzení, že Martialis psal skvělé epigramy. Zároveň doufá, na základě precedentů z minulosti, že jeho adresát, nepříliš zběhlý v této sféře, si jeho promluvu vyloží jako tvrzení, že Martialis psal skvělé epitafy, a že to z jeho reakce vyjde najevo: cílem je, dejme tomu, pobavit společnost na Karlův účet. V takovém případě je na místě říci, že Pavel si přeje být dezinterpretován, což znamená: Pavlova intence typu I_1 kontrastuje s jeho intencí typu I_2. Povaha jeho triku pak (triviálně) vylučuje, aby si přál naplnění Karlových intencí I_3 a I_4 – musí doufat v pravý opak. Důsledkem je, že se zhroutily korelace mezi podmínkami naplnění intencí z našeho výčtu – rozpadl se soubor ekvivalencí plynoucích bezprostředně z Davidsonova vymezení významu promluvy. To, jak je zřejmé, vůbec neznamená, že Pavlovy intence jsou vzájemně neslučitelné: Pavlova pozice je zcela koherentní a jeho promluva neztrácí schopnost nést zcela určitý význam (význam tvrzení, že Martialis psal skvělé epigramy).[24]

24 Intuitivní přijatelnost (ne-li samozřejmost) tohoto předpokladu může vzbudit dojem, že i ve sféře privátní konverzace je význam promluv určován výhradně konvenčně fixovanými pravidly, v nichž nepřipadá žádná role záměrům mluvčího a jejich rozpoznání adresátem. Karlův trik na ně zjevně spoléhá a má naději v tom uspět. Ale změňme scénář. Jsme svědky toho, jak Jana říká: „Havlíček psal vtipné epitafy." Z kontextu je všem zřejmé (přesně jako v kauze paní Malapropové),

Máme zde tedy příklad triku, v němž se intence mluvčího rozbíhají dvěma směry – jeden ke konvenčnímu, institucionálně vymezenému významu aktu, který se chystá vykonat, druhý k adresátově interpretaci téhož aktu, která je předjímána (a zamýšlena) jako neslučitelná s konvenčním významem. Nabízí se otázka, zda takový tah bude dávat smysl, když ho popíšeme v ne-intencionálním, čistě instrumentálním slovníku, který jsme si zvykli nasazovat při popisu výkonů AI. Pokud ano, zařadí se tím náš trik mezi tahy, které lze připisovat systémům AI nezávisle na tom, zda jim přiznáme intencionalitu.

9. Schopnost LLMs vykonat stejný tah na ne-intencionální bázi a přecházet mezi typy diskursu lišícími se ve způsobu konstituce významů promluv

Předpokládejme tedy systém AI nastavený na maximalizaci pozitivních reakcí na své výstupy (srov. výše kap. 6 a zmínku o RLHF v pozn. č. 17). To znamená, že ze všech možných (systému dostupných) tahů v jednotlivých komunikativních situacích se v jeho fungování prosadí ten, který s největší pravděpodobností vyvolá takové přijetí. Součástí jeho databáze, permanentně doplňované o nová data získaná z průběhu komunikace, je registrace faktu, že Karel, jeden z účastníků komunikace, opakovaně užívá výraz „epigram" v kontextech, v nichž se pravděpodobnost jeho výskytu blíží nule a jejichž nejpravděpodobnější slovní obsazení představuje výraz „epitaf" nebo „náhrobní nápis". A korelativně, Karel reaguje na užití slova „epigram" větami obsahujícími slova „epitaf" či „náhrobní nápis". Dále systém opakovaně zaznamenal, že odhalování takových diskrepancí sklízí u ostatních účastníků komunikace pozitivní reakce: reakce toho typu, na jejichž vyvolání je systém nastaven. Toto nastavení spolu s registrací zmíněného faktu o Karlovi povede k tomu, že systém navodí situaci, v níž se projeví Karlova deviace, a učiní tak užitím některé z vět obsahujících slovo „epigram".

Malá rekapitulace: navrhli jsme zvážit jazykový trik se slovem „epigram", abychom ukázali, jak Davidsonův radikální intencionalismus (paradoxně) redukuje intencionální stránku komunikace. Intencionální báze triku zahrnovala kontrast mezi zamýšleným významem promluvy odvozeným od konvenčního významu slova „epigram" a zamýšlenou interpretací. Mimokomunikativní motivací bylo navodit situaci, v níž se projeví jistý komunikační handicap adresáta. Pak jsme ukázali, jak se stejný trik realizuje ve výkonech

že Jana neměla na mysli náhrobní nápisy, ale (v této verzi) krátké satirické básně. Za těchto okolností je zcela přirozené přiznat, že se jí podařilo (i když poněkud nekonformní cestou) vykonat tvrzení, že Havlíček psal vtipné epigramy, v souladu s Davidsonovým pojetím významu promluvy a s jeho výkladem malapropismů (viz výše kap. 5 a 6).

systému AI postrádajících intencionální bázi: v těchto výkonech se dosahuje téhož účinku jako v lidské verzi využitím paralelní diskrepance (zde mezi adresátovou a statisticky pravděpodobnou slovní reakcí na výskyt výrazu „epigram“).

Zároveň bychom si měli připomenout, že systém AI obstál i v opačné situaci, o níž byla řeč výše. Šlo o případ paní Malapropové a jejího nestandardního užití slova „epitaph“. Jak měla ukázat naše úvaha v kap. 6, systém byl na stejné bázi jako v kauze Karel (tj. na základě vyhodnocení stupně pravděpodobnosti výskytu jistých slov v jistých kontextech) schopen najít standardní verzi nestandardní promluvy paní M. a založit na ní vlastní reakci na promluvu. Reakci, o níž bychom v případě lidského mluvčího s uznáním řekli, že respektuje skutečný komunikační záměr paní M., navzdory konvenčnímu významu slov, která zvolila. To ukazuje, že operace schopné překlenout kontrast mezi konvenčním významem promluvy a její nestandardní interpretací (případ Karel), a právě tak kontrast mezi nestandardním užitím slova a jeho konvenčním významem (případ Malapropová) jsou v možnostech systému AI, a to nezávisle na tom, zda mu přiznáme intence a schopnost operovat s pojmem jazykové konvence a komunikativní intence.[25]

To stačí k tomu, abychom systému připsali související schopnost: přecházet v průběhu komunikace mezi typem diskursu, v němž význam promluvy určuje aplikace jazykových konvencí na užitou větu,[26] a typem diskursu à la Davidson, v němž se významy promluv odvozují od rozpoznaných komunikativních intencí mluvčích. Schopnost přecházet mezi těmito typy diskursu zde znamená jak schopnost zaznamenat takový přechod a přizpůsobit se mu v reakcích na promluvy účastníků, tak schopnost iniciovat takový přechod (v situacích, v nichž to systém vyhodnotí jako statisticky nejčetněji zastoupený krok, případně krok schopný s vysokou pravděpodobností vyvolat pozitivní reakce).[27]

25 Něco jiného je schopnost operovat s *výrazy* jako „konvence“ a „intence“: tu lze u LLMs, trénovaných na rozsáhlých korpusech textů, předpokládat automaticky, míníme-li tím schopnost užívat je ve svých výstupech v obvyklých (očekávatelných) kontextech na standardních místech ve struktuře vět.

26 Pomíjím nezbytný příspěvek kontextu tam, kde ho vyžaduje sémantická funkce užitých výrazů – což je typicky případ indexických výrazů a indexických parametrů, jako je gramatický čas sloves.

27 Lze předpokládat, že stejný výsledek by (stejnou cestou) přinesla diskuse o schopnosti AI přecházet, v reakci na vývoj konverzace, mezi tzv. seriózním a fikčním diskursem, právě tak jako přizpůsobit své výkony přechodům mezi různými žánry a styly, od volné konverzace k vědeckým debatám či textům krásné literatury. K přechodům tohoto druhu v lidské komunikaci jsem se pokusil něco říci v článku: Fikční tvrzení, tvrzení o fikci a pohyb na rozhraní. *Filosofický časopis*, 70, 2022, č. 4, s. 673–700. K samotné tvorbě fikčních textů na bázi AI viz nejnověji Pioreckỳ, K. – Husárová, Z., *Kultura neuronových sítí. Syntetická literatura a umění (nejen) v českém a slovenském prostředí*. Praha, Ústav pro českou literaturu AV ČR 2024. Dostupné na:

10. Predikce místo závěru

V úvodu jsme si uložili zdrženlivost v několika směrech: jedním z nich byly aktuální spory o to, zda výkony AI v podobě produkce syntakticky správně utvořených vět můžeme klasifikovat jako mluvní akty typu tvrzení, otázek, slibů atd. Naše abstinence spočívala v důsledné snaze obejít se bez těchto termínů.[28] Debata, která se k nim váže, tím pro nás neztratila význam: na důkaz toho bych k ní teď rád řekl pár slov.

Omezme se zde na otázku, zda systémy AI typu LLMs jsou (v principu) schopny výkonů, jimž bychom mohli přiznat status tvrzení, a zkusme srovnat dva možné scénáře. Možností, která se bezprostředně nabízí autorům vstupujícím na toto pole z filosofie jazyka, je konfrontovat zaznamenatelné výkony AI s jejich preferovanými definicemi tvrzení. Protože v této druhé věci můžeme stěží doufat v obecný konsensus, máme zde slušné vyhlídky na debaty jdoucí ne snad *in infinitum*, ale *in indefinitum*.

Zvažme alternativní scénář. Předpokládejme, že někdy v blízké budoucnosti se AI stane reálnou a obecně rozšířenou součástí našeho každodenního života. To (mimo jiné) zahrnuje, že v rámci rutinní komunikativní praxe budeme AI adresovat své otázky, z opačné strany přijdou odpovědi v podobě syntakticky správně utvořených vět, budeme zvažovat jejich pravdivost či nepravdivost, příležitostně vyslovíme své námitky, reakcí budou argumenty ve prospěch původních odpovědí, kriticky zvážíme jejich relevanci a přesvědčivost, a samozřejmě nejen to: budeme AI zadávat praktické úlohy nejrůznějšího druhu a potom hodnotit jejich provedení jako adekvátní (anebo naopak – tak jako u lidí), AI nám občas navrhne dělat věci, které nám v dané situaci a vzhledem k našim preferencím budou dávat smysl (anebo ne – tak jako u lidí), někteří z nás se budou obracet k AI s nadějí na reakci ve vlídném a povzbudivém tónu (jíž se nedočkali od svých blízkých) a tak dále. Půjdou-li věci touto cestou,[29] pak v jistém okamžiku *nevyhnutelně* dospějí do fáze, kdy pro nás bude absurdní, svévolné a neslučitelné s naší jazykovou intuicí popírat, že vedeme regulérní komunikaci, v níž obě strany vykonávají tvr-

https://kramerius.lib.cas.cz/view/uuid:5c904eb4-12be-4d17-974c-46af8ca13de8; [cit. 16. 9. 2025]. Ke schopnosti takto vzniklých textů dosahovat čtenářských účinků nerozlišitelných od těch, jež navozují lidmi vytvořená literární díla, viz např. Hopkins, J. – Kiela, D., Automatically Generating Rhythmic Verse with Neural Networks, 2017. Dostupné na: https://research.fb.com/publications/automaticallygenerating-rhythmic-verse-with-neural-networks/; [cit. 1. 9. 2025].

28 Tak jako u termínů typu „přesvědčení" nebo „záměr", důvodem byla snaha ukázat, jak daleko lze s tímto omezením dojít při popisu a výkladu netriviálních operací, jako jsou přechody mezi různými typy diskursu nebo jazykové triky.

29 Na námitku, že dávno jdou touto cestou, a nejspíš na ní dospěly až k bodu, o němž se právě chystám mluvit, odpovídám „možná, ale dovolte mi zachovat si zdrženlivost z důvodu zmíněného v pozn. č. 28".

zení, kladou otázky, vzájemně se vyzývají k nejrůznějším akcím atd. A zcela rutinně budeme přistupovat k výkonům AI jako k aktům tohoto druhu, aniž bychom se zajímali o povahu procesů v pozadí těchto výkonů (na úrovni umělých neuronových sítí) – tak jako se nestaráme o procesy na úrovni neuronových spojů v mozcích našich bližních, kterým připisujeme tvrzení, otázky, sliby atd.

Dojde-li k tomu, důsledky budou zásadní. Výrazy jako „tvrzení", „otázka", „prosba" atd. nejsou (na rozdíl od slov jako „ novičok", „covid" nebo „artritida") teoretické termíny, které by podléhaly dělbě jazykové práce, v níž se podrobujeme autoritě expertů, spoléháme se na jejich úzus (na němž se nepodílíme), případně na jejich definice (které zpravidla neznáme) a jsme připraveni přijmout z této strany korekce našich laických užití.[30] Jinými slovy, významy a s nimi podmínky aplikability slov jako „tvrzení" se *beze zbytku* vymezují v každodenní komunikativní praxi. Pokud tedy nastane stav věcí popsaný v předchozím odstavci, případný protest filosofů či lingvistů proti připisování aktů jako tvrzení systémům AI bude *de facto* protest proti přirozenému jazyku.[31]

Nic z toho, co bylo právě řečeno, nezpochybňuje ani nezužuje prostor pro filosofické teorie mluvních aktů. Tak jako dosud bude stále platit, že tyto teorie mohou odvést cennou práci na poli analýzy a výkladu fungování každodenní jazykové komunikace – v daném případě zejména identifikace kritérií, která se v této sféře reálně uplatňují při určování výpovědní funkce či síly („illocutionary force") promluv a ve volbě našich reakcí na ně. Teorii, která by se v této věci ocitla v rozporu s komunikační praxí, například by vyústila v definici tvrzení, jejíž kritéria (nutné a ve svém souhrnu dostačující podmínky vymezené v definiens) by se rozcházela s tím, jak reální interpreti klasifikují promluvy reálných mluvčích, by tak zbývala jediná možnost: připustit, že se míjí se svým předmětem, a začít znovu na jiné bázi.

Jiná věc je, zda obecně sdílená praxe vztahování slova „tvrzení" i na výkony AI by nutně znamenala změnu dosavadního významu tohoto slova. Jak konstatoval Ludwig Wittgenstein (ve svých úvahách o jednání podle pravidla), význam slova neurčuje předem možnosti (a limity) jeho užití ve všech myslitelných budoucích situacích (není to mechanismus, v němž jsou pře-

30 Klasický výklad pojmu dělby jazykové práce podal (v rámci své argumentace pro sémantický externalismus) Hilary Putnam ve stati: The Meaning of „Meaning". In: týž, *Mind, Language and Reality*. New York, Cambridge University Press 1975.

31 Čtenáři *Filosofických zkoumání* pak nejspíš pocítí nutkání komentovat to slovy, že "… philosophical problems arise when language goes on holiday" (Wittgenstein L., *Philosophical Investigations*. Oxford, Blackwell 1953, § 38). A připomenou, že "Philosophy may in no way interfere with the actual use of language; it can in the end only describe it. For it cannot give it any foundation either. It leaves everything as it is." (tamtéž, § 124).

dem „obsaženy“ všechny úkony, které může vykonávat[32]): v nových kontextech se významy (a s nimi podmínky aplikability) slov „dourčují“ v reálných praktikách, které se ustálí v jazykovém společenství. V našem případě by šlo o změnu významu jen tehdy, kdyby aktuální způsob užívání slova „tvrzení“ (a s ním spojený způsob, jakým reagujeme na promluvy jiných) vylučoval aplikaci tohoto slova na výkony AI. Řekl bych, že současná praxe je v této věci neutrální.[33]

32 Tamtéž, § 193. Právě tak nefunguje jako koleje, které určují předem jedinou možnou dráhu našeho pohybu (tamtéž, § 218).

33 K prvotním zdrojům inspirace, z nichž vzešla tato stať, patřily články a přednášky kolegů, kteří dříve (a razantněji) než já vstoupili do diskusí o AI: Víta Gvoždiaka, Vladimíra Havlíka, Tomáše Hříbka, Juraje Hvoreckého a Jakuba Mihálika. Prvnímu z nich navíc vděčím za odkazy k textům, s nimiž bych se bez něj nejspíš minul. Za cenné podněty děkuji autorům/autorkám anonymních recenzních posudků *Filosofického časopisu*: na část z nich reaguji v předchozím textu (viz pozn. č. 9 a 18).