Nový pohled na výpočty a umělá inteligence

Jiří Wiedermann

Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha, Česká republika
a
Český institut informatiky, robotiky a kybernetiky
ČVUT Praha, Česká republika
Email: jiri.wiedermann@cs.cas.cz

Abstrakt

Klasický pohled na výpočty je chápe jako procesy generované počítačem, tj. soustřeďuje se na to, JAK jsou výpočty realizované. Nový pohled se zajímá o to, CO výpočty dělají, co je jejich smyslem — a to je generování znalostí. Výpočty tudíž chápeme jako procesy, které generují znalosti nad danou znalostní doménou v rámci příslušné znalostní teorie. Inteligence je pak schopnost získávat informace a transformovat je na znalosti, které jsou dále využívány pro řešení problémů. Ukážeme, že toto pojetí výpočtů umožní přirozeným způsobem definovat některé vyšší kognitivní funkce jako je zdůvodňování svých akcí, sebe-uvědomění, introspekce, porozumění znalostem, svobodná vůle, kreativita a také porozumění algoritmickým mechanismům, které stojí za rozvojem inteligence. Tento pohled je přínosný tím, že na elementárním abstraktním modelu kognitivního systému, který není zatížen žádnými technickými detaily, ukazuje, že všechny výše zmíněné kognitivní funkce souvisejí se specifickými znalostmi, které jsou generované v rámci téhož modelu.

1 Úvod

Výpočty jsou tradičně chápány jako procesy, které probíhají v počítačích. Pokud bychom chtěli tuto definici výpočtu zpřesnit, musím zpřesnit pojem počítače. Za tím účelem se v teorii obyčejně uvažuje Turingův stroj, o kterém je známo, že tento model v principu zachycuje výpočetní schopnosti velké třídy současných digitálních počítačů. Nicméně, v praxi, a zejména v umělé inteligenci nebo v biologii, často považujeme za výpočty i procesy, které probíhají nikoliv v lidmi uměle zhotovenými zařízeními (tj. v počítačích), ale také např. v mozku lidí či zvířat, v rostlinách anebo buňkách. Ve fyzice se vyskytují úvahy, že dokonce na celý vesmír lze nahlížet jako na obrovský počítač, který se řídí „instrukcemi“, které představují fyzikální zákony. Takže pokud se chceme dozvědět něco o výpočtech, musíme zkoumat schopnosti a meze těchto zařízení. To znamená, že se soustředíme na problém, JAK jsou výpočty realizované. Takový pohled je sice technicky zajímavý, avšak vede na (klasickou) teorii výpočtů, která je strojově závislá a nemá potenciál zachytit, jaký vlastně je smysl, cíl počítání — co vlastně výpočty „dělají“ či mohou „dělat“, ať už pro nás, pro lidi, anebo pro to zařízení, které výpočet realizuje.

Pokud se tedy ptáme, co výpočty vlastně dělají, tak jedinou smysluplnou odpovědí je, že produkují znalosti. Toto je výchozí bod tzv. *znalostní teorie výpočtů,* která vychází z prací autorů van Leeuwena a Wiedermanna (2013, 2014, 2015a, 2015b, 2017). Dle této teorie jsou výpočty chápány jako procesy, které generují znalosti nad danou znalostní doménou v rámci příslušné znalostní teorie. V našem příspěvku ukážeme, že tento pohled má velký potenciál zejména pro umělou inteligenci, protože umožňuje přirozeným způsobem definovat (a chápat) netriviální kognitivní funkce jako je zdůvodňování svých akcí, sebe-uvědomění, introspekce, porozumění znalostem, svobodné vůli a kreativitě a přináší vhled do algoritmických mechanismů, které stojí za rozvojem inteligence. Toto lze považovat za zásadní přínos k problematice výpočetních kognitivních systémů, protože jiné známé přístupy k popisu těchto funkcí používají daleko složitější modely kognitivních systémů než je náš znalostní model. Tyto jsou obyčejně tak složité, že výsledný popis či definice kognitivních funkcí opět mají charakter, který spíše záleží na vlastnostech daného modelu, na jeho architektuře nežli na obecných vlastnostech uvažovaných funkcí. Podrobný přehled realizovaných kognitivních architektur lze nalézt v práci a v on-line katalogu Samsonoviche (2010).

Struktura příspěvku je následující. V části 2 uvedeme základní ideje znalostního přístupu k výpočtům. V části 3 vysvětlíme, jak nový pohled na výpočty přirozeným způsobem vede k definici základních vyšších kognitivních funkcí, jakými jsou důvodění, sebeuvědomění, introspekce, porozumění, svobodná vůle, kreativita a sebe-zdokonalování znalostních teorií. Poslední část 4 shrnuje základní poznatky našeho příspěvku.

2 Výpočet jako generování znalostí

V souladu se znalostní teorii výpočtů budeme nahlížet na znalosti jako na výsledek nějakého výpočetního procesu, který pracuje nad jistou znalostní doménou tak, že kombinuje její prvky — *elementární znalosti* — do odvozených, často složitějších konstrukcí, které tvoří novou znalost, opět nad danou doménou. Pro kombinaci těchto prvků používá výpočet množinu (odvozovacích) pravidel, která může být předem daná, anebo se může tvořit pomocí učení během velkého počtu různých výpočtů nad danou doménou. Inteligentní systém tímto způsobem pracuje s více či méně formální teorii, která zachycuje vlastnosti dané znalostní domény a způsoby odvozování nových znalostí, stále v rámci dané domény.

Jak jsme již zmínili, přitom nás nebude v první řadě zajímat, JAK daný výpočet probíhá, ale CO výpočet počítá — jaká znalost je generována v průběhu výpočtu. Pod tímto zorným úhlem se stává schopnost generovat znalosti poznávacím kritériem těch procesů, které budeme nazývat jako výpočetní procesy neboli výpočty. Inteligentní systémy jsou tudíž speciálními případy výpočetních systémů, u kterých je schopnost generovat znalost maximalizována v tom smyslu, že takové systémy jsou schopny generovat znalost nad libovolnými znalostními doménami modelujícími velké části reálného světa anebo různých věd. To je v protikladu s praxí současných AI systémů, které jsou zpravidla specializovány na specifické, většinou značně omezené znalostní domény. Procesy, které negenerují znalost, nebudeme považovat za výpočty. Zde narážíme na problém — jak poznáme, co je znalost? To je fundamentální filozofická otázka, na kterou filozofové doposud nenalezli odpověď. Jedno je jisté — co se jednomu jeví jako znalost, pro druhého to může být buď samozřejmost, anebo to nemusí vůbec považovat za znalost. Pojem znalosti není tedy absolutní, ale je *závislý na pozorovateli* (observer dependent). Závisí tedy na tom, co již pozorovatel zná. Ve znalostním přístupu se to řeší tak, že znalost se definuje v rámci nějaké *znalostní domény,* nad kterou výpočet operuje. Všechny znalosti o nějaké podmnožině znalostní domény jsou zachyceny pomocí *znalostní teorie*, která může být více či méně formální, anebo zcela neformální. V rámci této teorie popisují *axiomy* elementární znalosti, které odpovídají (reprezentacím) objektů ve znalostní doméně a jejím vlastnostem. Způsoby, jak lze z takových elementárních znalostí konstruovat nové, odvozené znalosti jsou popsány pomocí *odvozovacích pravidel.* Výpočetní procesy jsou svázány s odpovídající znalostní doménou prostřednictvím znalostní teorie, se kterou výpočty přímo či nepřímo pracují, pomocí následující podmínky: *cokoliv lze odvodit v rámci dané teorie musí být podporováno příslušným výpočetním procesem.* Pokud je tomu tak, pak to, jaká znalost může anebo nemůže být generována nad danou znalostní doménou, a „kvalita“ takto generované znalosti (tj. např. její shoda s pozorováním) závisí výlučně na vlastnostech odpovídající znalostní teorie.

Všimněme si, že znalostní přístup k výpočtům je *strojově nezávislý,* protože platí pro jakýkoliv proces realizující odvozování v rámci dané znalostní teorie. Taktéž je *algoritmicky nezávislý,* protože se nezajímáme o to, jakými postupy je výpočetní proces realizován. V neposlední řadě je také *nezávislý na reprezentaci,* protože v našem přístupu nepředpokládáme žádnou speciální reprezentaci znalostí.

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.* 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 argumentací, která lze zachytit v přirozeném jazyce. V typickém případě mají popisné domény rozsáhlé znalostní báze (jako např. obsah internetu) a relativně krátké odvozovací řetězce.

Význačným příkladem inteligentního systému vykazujícím lidskou inteligenci je mozek společně s přirozeným jazykem. Mozek umožňuje odvozovací procesy v neformální teorii, kterou lze popsat v přirozeném jazyce. Samozřejmě, že v principu můžeme namísto mozku uvažovat jakýkoliv počítač s podobnými vlastnostmi, i takový, o kterém doposud nevíme, že existuje; výsledkem bude systém umělé inteligence na lidské úrovni. Skutečnost, že znalostní přístup k výpočtům umožňuje pracovat i s takovými nedokonale definovanými pojmy je předností našeho modelování. To nám umožní dosáhnout nové porozumění problematice generování znalostí, které zatím nebylo dosaženo jiným způsobem. Pro formalizaci tohoto přístupu viz práci van Leeuwena a Wiedermanna (2017). Přehled dosavadních výsledků z oblasti kognitivních výpočtů lze nalézt v práci Wiedermanna a van Leeuwena (2015a).

3 Kognitivní funkce jako nadstavba nad mechanismem generováním znalostí

Vraťme se teď zpět k podmínce, která svazovala výpočetní proces se znalostmi, které lze nad danou znalostní doménou odvodit. Zde jsme požadovali, aby jakákoliv znalost, kterou lze odvodit (formou důkazu v dané znalostní teorii nad danou znalostní doménou) byla prokazatelně (tj. opět: musí existovat důkaz) podporovaná příslušným výpočetním procesem.

Pokud umělý kognitivní systém pracuje tímto způsobem, může při vhodné organizaci své činnosti realizovat některé netriviální vyšší kognitivní funkce, jejichž definice není na první pohled zřejmá, a tím spíše

jejich realizace.

Důvodění (accountability) Důvodění znamená schopnost zdůvodnění svých rozhodnutí, tj. kognitivní agent může na vyžádání podat zdůvodnění svých akcí, vysvětlení „jak na to přišel“, viz Kroll a spol. (2016). Za tím účelem stačí, aby agent společně s výslednou znalostí generoval i důkaz, který použil při jejím odvozování. Tento důkaz již samozřejmě může být prezentován ve formalismu, kterému uživatel rozumí. Takový důkaz umožní uživateli „kontrolu správnosti“ výsledku v rámci odpovídající znalostní teorie v porovnání s tím, co uživatel očekával na základě svého zadání.

Sebeuvědomění (awareness) Díky schopnosti důvodění znalostní systém „ví“ (má informaci), jaký problém řeší a je schopen podat vysvětlení, jak ho řeší. Znalost o sobě může být součástí znalostní teorie, která řídí činnost systému.

Introspekce Na základě skutečnosti, že si kognitivní systém pamatuje své předchozí řešené úkoly a způsoby jejich vyřešení, může se k ním vracet, znovu je podrobit zkoumání a využít je při řešení nových problémů pomocí analogie anebo vylepšit jejich původní řešení s ohledem na nové poznatky, které systém mezitím mohl získat.

Znalostní porozumění Schopnost důvodění, sebeuvědomění a introspekce dohromady představují schopnost porozumění znalostní doméně, s kterou systém pracuje. Systém je schopen vysvětlit význam termínů, které používá a na základě své předchozí zkušenosti (které jsou zapamatovány v příslušné znalostní bázi) kreativně je aplikovat v novém kontextu. Pro plné porozumění reálnému světu je potřebné uvažovat vtělené systémy.

Svobodná vůle Budeme říkat, že kognitivní systém A má svobodnou vůli právě tehdy, když neexistuje kognitivní systém B, který výlučně na základě pozorování chování systému A v různých situacích dovede vždy předpovědět chování systému A v dané situaci. Tato naše definice se liší od standardních definicí (kterých je nepřeberné množství — viz např. odpovídající heslo na Wikipedii) a které vidí svobodnou vůli jakoby z vnitřního pohled systému, tj. subjektivně. Např. „svobodná vůle je schopnost zvolit si různé možnosti chování v dané situaci,“ anebo „schopnost chovat se ve výsledku jinak než na základě minulých událostí“. O tom, že si kognitivní systém zvolil z různých možností chování, anebo že se chystá zachovat jinak, než předtím, má informaci pouze systém samotný. A protože v obecném případě systému „nevidíme do hlavy“, nejsme schopni rozhodnout, jestli má svobodnou vůli. Je však zřejmé, že oba právě zmíněné případy (a další) zachycuje naše definice tak, že vnější pozorovatel není schopen predikovat budoucí akce systému. To znamená, že naše definice činí pojem svobodné vůle závislým na pozorovateli. Má však výhodu v tom, že činí problém, jestli má systém svobodnou vůli, rozhodnutelným (samozřejmě vzhledem k pozorovateli).

Kreativita Kreativita je projevem kreativního procesu, což je každý proces, který generuje znalost řešící problém, jenž je pro kognitivní systém nový. Je to protiklad rutinního procesu řešícího známý problém pomocí již známých postupů. Obecně, řešení problému vyžaduje najít znalost, která splňuje předem danou množinou podmínek. Jinými slovy, hledáme explicitní znalost, která je implicitně zadaná pomocí vlastností, které musí tato znalost splňovat. Pokud tento úkol modelujeme pomocí znalostního přístupu k počítání, tak výchozím postupem pro hledání řešení zadaného úkolu je použití „hrubé síly“ — systematického generování všech znalosti, které lze v systému (tj. v příslušné znalostní teorii) odvodit a zkoušení, jestli daná znalost nesplňuje zadané podmínky. To vypadá jako hrubě neefektivní až naivní přístup, ale zdá se, že tento přístup je v pozadí jakéhokoliv kreativního procesu. Jedná se tedy o speciální případ *objevování znalostí (knowledge discovery)*. Naštěstí, opakovaným používáním takového zpočátku neefektivního, leč univerzálního postupu jej lze kultivovat. Na objevování znalostí hledíme jako na jeden interaktivní, evoluční a potenciálně nekonečný sebezdokonalující se a učící se proces, jehož cílem je zlepšovat své vlastní kreativní schopnosti. Kultivaci procesu objevování znalostí lze popdle Wiedermanna a van Leeuwena (2015b) dosahovat zejména

- iterativním zjemňováním a/anebo rozšiřováním vyhledávacích kritérií na základě předchozích zkušeností anebo málo úspěšných pokusů;
- automatickou extrakcí a modifikací uživatelských kritérií, které zužují výběr možností na základě odpozorovaných preferencí užívatele a jeho emocí a prožitků (podobně, jako to dělá Google+). Toto se děje při každé příležitosti (tj. nejen pro řešení každého konkrétního problému zvlášť) a získané poznatky jsou využívány pro usměrňování jakéhokoliv objevného procesu a pro uspořádávání objevovaných výsledků.
- využíváním předchozích zkušeností, které mohou být odvozeny z epizodické paměti;
- řízenou interakcí s okolím zaměřenou na získání dalších znalostí, které mohou být nápomocny při řešení konkrétního problému (podobně, jako když na Internetu vyhledáváme dodatečné informace).

Použitím takových kultivačních procedur se mění (doplňuje) jak systémová báze znalostí, tak i mechanismy práce s ní.

Sebe-zdokonalování znalostních teorií Samotnou schopnost odvozovat další znalosti v rámci dané znalostní teorie však nelze považovat za hlavní znak inteligentních systémů. Tím je až jejich schopnost vylepšovat svoji znalostní teorii, pomocí a v rámci které

tyto systémy generují své znalosti. Inteligence takových systémů prokazatelně roste, protože získávají kvalitativně nové znalosti o znalostní doméně, nad kterou pracují.

Při použití kultivačních procedur popsaných v předchozím případě se může stát, že systém získá novou znalost, které je ve sporu se znalostmi, které již systém má. Buď si takovou znalost odvodí systém sám, anebo ji získá „zvnějšku“ (třeba z Internetu) anebo vlastním pozorováním a odhalením nesouladu pozorování s vlastní teorii. Pro takové případy musí mít vyspělý systém mechanismy, které odhalí logickou inkonsistenci své znalostní teorie. Takovou vadu lze odstranit jedině změnou příslušné teorie. Systémy, které jsou schopny sebe-zdokonalovat svoji znalostní teorii, a jsou navrženy tak, aby systematicky vyhledávaly rozpory své teorie s fakty, zřejmě mohou automaticky, pomocí svých vlastních mechanismů, zvyšovat svoji vlastní inteligenci, při jakékoliv rozumné definici pojmu inteligence. Takové sebe-zdokonalování může pokračovat do té doby, pokud existují rozporuplná fakta a systém je objeví, a také pokud ve znalostní doméně existují nové, doposud neprozkoumané objekty a jevy (Bostrom 2014). Takové systémy pak mohou, alespoň teoreticky a v některých směrech, překonat inteligenci lidskou (Wiedermann a van Leeuwen 2017). Pojem sebe-zdokonalovacích znalostních teorií překonává svým dopadem dosavadní obecné představy o tzv. sebezdokonalovacím software (viz např. Bostrom 2014), protože identifikuje jako nutnou podmínku pro nárůst inteligence kognitivních systémů znalostní data, jejich kvalitu i kvantitu (a tedy — kvalitu příslušných znalostních teorií) a nikoliv efektivitu odpovídajících odvozovacích mechanizmů.

4 Závěr

V příspěvku jsme ukázali nový pohled na výpočty, který je chápe jako procesy generující znalosti a využívá tento pohled k definici a pochopení netriviálních vyšších kognitivních funkcí. Takovými funkcemi jsou důvodění, sebeuvědomění, introspekce, porozumění, svobodná vůle, kreativita a sebe-zdokonalování znalostních teorií. Tyto funkce nelze elegantně popsat pomocí klasického pohledu na výpočty, který považuje za výpočty jakékoliv, i nesmyslné procesy, generované různými modely počítačů. Takový pohled je nutně strojově orientovaný a tudíž neposkytuje dostatečně abstraktní a obecný rámec pro definici zmiňovaných kognitivních funkcí a jejich pochopení. Naopak, náš pohled přes teorii kognitivních výpočtů ukazuje pomocí elementárního abstraktního modelu kognitivního systému, který není zatížen žádnými technickými detaily, že všechny výše zmíněné kognitivní funkce souvisejí se specifickými znalostmi, které jsou generované v rámci téhož modelu.

Poděkování

Tento příspěvek vznikl za částečné podpory GA ČR v rámci grantové úlohy 15-04960S, institucionálního plánu ÚI AV ČR RVO 67985807 a programu Strategie AV21.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press

Kroll, J.A. a spol. (2016). Accountable Algorithms (March 2, 2016). University of Pennsylvania Law Review, Vol. 165, 2017 Forthcoming; Fordham Law Legal Studies Research Paper No. 2765268. Available at SSRN: https://ssrn.com/abstract=2765268

Samsonovich, A.V. (2010). Toward a Unified Catalog of Implemented Cognitive Architectures. BICA 221 (2010): 195-244, on-line katalog http://bicasociety.org/cogarch/.

van Leeuwen, J. a Wiedermann, J. (2017). Knowledge, representation and the dynamics of computation. In: G. Dodig-Crnkovic, R. Giovagnoli (Eds): *Representation and Reality: Humans, Animals and Machines.* Berlin: Springer

Wiedermann, J. a van Leeuwen, J.(2013). Rethinking computation. In: *Proc. 6th AISB Symp. on Computing and Philosophy: The Scandal of Computation - What is Computation?*, AISB Convention 2013 (Exeter, UK), AISB, pp. 6-10

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, s. 356-363