Jan Veselý
Fakulta filozofická Západočeské univerzity v Plzni; veselyh@ff.zcu.cz

Cameron J. Buckner: From Deep Learning to Rational Machines

What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence

New York, Oxford University Press 2024. 438 s.

Deklarovaným hlavním záměrem titulu *From Deep Learning to Rational Machines* je prozkoumat vztah mezi umělou inteligencí, hlubokým strojovým učením a empiristickými teoriemi kognice. Jeho autorem je Cameron Buckner, který působí jako profesor filosofie na *Donald F. Cronin Endowed Chair* na *University of Florida*, kde se zaměřuje na výzkum non-lidské kognice. Kniha je mezioborovou publikací, která osciluje na pomezí AI, filosofie a kognitivní vědy.

Buckner v knize prezentuje koncepci *Doménově-obecné modulární architektury* (DoGMA). Ústřední tezí titulu je, že tato koncepce, vycházející z principů umírněného empirismu, nabízí vhodný teoretický rámec pro rozvoj systémů umělé inteligence. Zajímavé je, že DoGMA je inspirovaná empiristickou psychologií mohutností (*faculty psychology*), které zahrnují percepci, paměť, imaginaci, pozornost a sociální kognici. Tyto mohutnosti Buckner chápe jako mechanismy či „funkcionality", které jsou schopné zpracovávat širokou škálu dat napříč různými oblastmi kognice.

V rámci DoGMA každé z těchto mohutností odpovídá samostatný modul, který je realizovaný hlubokou neuronovou sítí. Lze říci, že tímto je na architektonické rovině modelována struktura lidské kognice. Moduly ovšem nejsou zcela nezávislé a informačně uzavřené.[1] Vzájemně spolu interagují, soutěží o zdroje a dynamicky zachycují lidskou racionalitu i na procesuální rovině.

Každá mohutnost je přiblížena na příkladu jejího pojetí u vlivného empiristického filosofa.[2] Mechanismy, které jsou pro tuto mohutnost charakteristické, jsou následně porovnávány s mechanismy, které zajišťují chod konkrétních architektur strojového učení. Kniha má celkem sedm kapitol.

1 Bucknerovské moduly nejsou – na rozdíl od fodorovských modulů – doménově specifické a informačně uzavřené. Viz Buckner, C. J., *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artifical Intelligence*. New York, Oxford University Press 2024. Srov. s vlivným pojetím modularity J. Fodora, který zastává opačnou pozici: Fodor, J. A., *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MIT Press 1983.

2 Adjektivum „empiristický" je chápáno v širokém slova smyslu a označuje britské empiristy (Locka, Huma), ale též další myslitele z historie filosofie, kteří tíhnou k empiristickému pojetí kognice.

1. V první kapitole Buckner prezentuje diskusi mezi nativismem a empirismem v kognitivních vědách a AI i v historii filosofie. Tuto diskusi, která má kořeny již v antické filosofii, přitom nově reinterpretuje jako spor mezi zastánci „doménově obecného" a „doménově specifického" přístupu k mysli. Jako doménově specifické lze dle Bucknera chápat takové kognitivní struktury, které jsou vrozené a vysoce specializované (např. vrozené pojmy, vrozený modul pro gramatiku apod.). Doménově obecné schopnosti naopak zhruba odpovídají pojetí mentálních mohutností u významných empiristických filosofů.

Buckner využívá tento spor k tomu, aby odstínil svou vlastní pozici od (i) radikálního empirismu a (ii) radikálního nativismu.

(i) Radikální empirismus tvrdí, že mysl je „nepopsaná tabule" (tabula rasa), do které se pasivně otiskují vnější podněty. Odmítá vrozené reprezentace a vrozené mechanismy (včetně doménově-obecných mohutností). Jediné, co je v tomto pojetí připuštěno, je hrstka jednoduchých asociativních principů. Buckner tuto pozici odmítá jako extrémní a jen obtížně udržitelnou. Podobně kritický je i k radikálnímu nativismu (ii), který naopak vrozené reprezentace a specializované vrozené mechanismy (tj. doménově specifické struktury) postuluje.

Pozice (i) a (ii) tedy Buckner zavrhuje a propaguje (iii) „kompromisní" řešení sporu spočívající v přijetí umírněného empirismu. V jeho rámci jsou odmítnuty vrozené ideje a vrozené (doménově-specifické) struktury. Doménově-obecné psychologické mohutnosti jsou však naopak předpokládány.[3]

Podle Bucknera byla řadě empiristických filosofů předchozích epoch mylně přisuzována silnější pozice (ii), i když ve skutečnosti zastávali pouze slabší pozici (iii). Tato skutečnost dle něj mohla vést nativisty k přehnaně příkré kritice empiristických pojetí kognice, která se ve své umírněné interpretaci dle Buckenera ukazují pro modelování lidské racionality jako vhodná.

Buckner se snaží v knize ukázat mj. i to, že moderní systémy hlubokého strojového učení alespoň částečně potvrzují teze přítomné v těchto klasických empiristických teoriích. I když však tyto systémy aproximují určité aspekty lidské kognice, dle Bucknera při modelování lidské racionality také stále vykazují řadu limitů a nedostatků.

2. Ve druhé kapitole Buckner představuje mj. metodologické principy pro hodnocení AI, kterými navrhuje nahradit stávající metriky hodnocení racionality u systémů umělé inteligence.

3. Zřejmě nejzajímavější částí knihy je její třetí kapitola. Ta obsahuje nejsilnější empirickou podporu Bucknerovy teze, že Doménově obecnou modulární architektu-

3 Minimálně na způsob „teoretického postulátu", resp. jakési užitečné fikce.

ru – respektive alespoň některé její aspekty – lze realizovat (i v aktuálně existujících) systémech hlubokého strojového učení.[4]

Buckner v této kapitole prezentuje Lockovu empiristickou teorii percepce a zejména zrakové perceptuální abstrakce, přičemž zkoumá její vztah k hlubokým konvolučním neuronovým sítím (DCNN). Perceptuální abstrakci u Locka Buckner interpretuje jako hierarchický proces, který postupuje od vstupních dat dodaných smysly až k obecným, vysoce abstraktním kategoriálním reprezentacím.

Podle Locka mají veškeré znalosti původ ve smyslech. Mysl díky nim získává mentální reprezentace vnějších předmětů. Ty mají podobu partikulárních idejí, které reprezentují konkrétní předměty nebo aspekty vnějšího světa v konkrétních situacích či stavech. Z partikulárních idejí jsou postupně abstrahovány rysy vyšších a vyšších úrovní obecnosti. Například z konkrétních partikulárních vizuálních reprezentací stromu, viděného postupně z perspektivy 1, perspektivy 2 až z perspektivy N, jsou abstrahovány takové rysy stromu, které umožní podřazení těchto partikulárních idejí pod obecnou ideu (pojem) stromu.

Podle Bucknera tedy neplatí, že se podněty do lockovsky chápané mysli pouze pasivně otiskují, podobně jako do prázdné voskové tabulky. V duchu umírněného empirismu naopak percepci vysvětluje jako aktivní a hierarchicky organizovaný kognitivní proces. Zajímavé je, že rozlišuje čtyři typy abstrakce, které chápe jako vzájemně komplementární a potenciálně implementovatelné v systémech strojového učení.

Při úlohách zaměřených na rozpoznávání obrazu podle Bucknera konvoluční neuronové sítě postupují způsobem, který je Lockovu pojetí perceptuální abstrakce analogický. DCNN ze vstupních dat postupně extrahují relevantní údaje (*features*) o předmětech. První vrstvy detekují typicky jednoduché vizuální rysy, například okraje a textury. Střední vrstvy DCNN konstruují složitější reprezentace, které odpovídají např. částem předmětů, a nejvyšší vrstvy následně klasifikují komplexní objekty jakožto jednotné celky.

Mezi Lockovým pojetím percepce, abstrakce a DCNN však kromě strukturálních podobností existují i rozdíly.

DCNN nejsou při rozpoznávání objektů podobně flexibilní jako lidé. Problematická pro ně je zejména omezená schopnost generalizovat naučené znalosti a používat je mimo původní kontext. Jinak řečeno, konvoluční sítě jsou omezené na data (či typ dat), která jsou obsažená v trénovacím data-setu, a příliš se jim nedaří využívat naučené znalosti na data jiného typu. (Jedná se o problém transferu znalostí.)

Lidé jsou – na rozdíl od DCNN – zobecňování mimo původní kontext schopní provádět bez větších problémů už po setkání s několika příklady a dokáží snad-

4 Viz Buckner, J., *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artifical Intelligence*, c.d., s. 203.

no integrovat podněty přicházející z více smyslových modalit. Jsou také schopní provádět tzv. kauzální inferenci, tedy mj. rozpoznat vztah mezi příčinou a účinkem a odlišit jej od pouhé statistické korelace. DCNN se naproti tomu spoléhají na prosté korelace mezi daty a kauzální vztahy nedokáží od pouhých korelací odlišit.

Buckner upozorňuje rovněž na další neduh konvolučních sítí, kterým je zranitelnost v situacích, kdy jsou vystaveny tzv. *adversariálním útokům*: Pokud je v rámci útoku cíleně provedena malá změna u několika pixelů v obrázku (na vstupu DCNN), kterou člověk typicky vůbec není schopný postřehnout, může tato změna vyústit ve zcela mylnou klasifikaci daného obrázku konvoluční sítí. Tento problém poukazuje na odlišnost ve zpracování senzorických informací u lidí a DCNN, i přesto, že oba typy percepce mají (mít) společné strukturální rysy. Neuronové sítě jsou sice inspirované strukturou vizuální kůry v mozku, nicméně je diskutabilní, do jaké míry jsou principy jejich fungování biologicky (a obecně empiricky) plausibilní.

4. Další mohutností, kterou Buckner zkoumá, je paměť. Tuto mohutnost přibližuje prizmatem Ibn Sínovy (Avicennovy) teorie, kterou usouvztažňuje se systémy založenými na posilovaném učení RL (*reinforcement learning*).

Avicenna předpokládá, že mohutnost paměti je v lidech realizována sadou mozkových komor, které vykazují hierarchickou organizaci. V těchto komorách dle něj dochází k postupnému zpracování informací přijatých prostřednictvím smyslů. Z dnešního pohledu lze chápat tuto teorii paměti jako překonanou a biologicky neplausibilní. Buckner si však všímá, že Avicennova koncepce paměti zahrnuje mj. také popis procesu hierarchické integrace podnětů, v jehož rámci je nová informace přijatá smysly postupně asimilována do širšího znalostního rámce. V tomto procesu hraje důležitou roli mj. i „záměr" agenta a emocionální či afektivní náboj, který pro něj daný podnět má. Buckner argumentuje, že na obecné rovině (např. hierarchie, afektivita) je tento přístup pro vývoj AI relevantní a inspirativní, a to i přesto, že Avicennova teorie obsahuje některé problematické a zastaralé biologické předpoklady.

Systémy fungující na základě posilovaného učení (RL) dokáží optimalizovat své rozhodování na základě zkušenosti a získané zpětné vazby, nicméně nedisponují strukturovanou dlouhodobou pamětí. Buckner upozorňuje, že tyto modely trpí vážným nedostatkem, tzv. *katastrofickou ztrátou paměti* (*catastrophic forgetting*). Učení nových informací u nich typicky vede k přepsání či „zapomenutí" dříve naučených znalostí. Díky tomu jsou RL neefektivní při kumulativním učení. Z tohoto důvodu je lze využívat pouze na jeden typ úloh, a to ten, na který byly natrénovány. V tomto ohledu tedy nedokáží modelovat lidskou racionalitu, jelikož lidé jsou na rozdíl od nich schopní vykonávat celou řadu úloh bez *katastrofické* úrovně *zapomínání* dříve natrénovaných dovedností nebo osvojených znalostí.

Buckner se domnívá, že problému *katastrofického zapomínání* se lze vyhnout, pokud systémy RL obohatíme o mohutnost paměti. Ta by měla být schopná – po-

dobně jako v Avicennově pojetí – hierarchicky zpracovávat a integrovat nové znalosti, a umožnit tak systémům RL osvojení širší škály dovedností. Tím, že paměť umožní uchování minulých zkušeností, může zároveň potenciálně umožnit aplikaci těchto znalostí napříč různými situacemi a kontexty. Do procesu zapamatování by také měla vstupovat vyhodnocená důležitost podnětu, resp. jeho afektivní náboj. Kromě dlouhodobé paměti Buckner uvažuje také např. o paměti krátkodobé nebo paměti epizodické.

5. Mohutnost imaginace Buckner představuje prizmatem pojetí této mohutnosti ve filosofii Davida Huma. Imaginaci přitom chápe jednak jako schopnost vytvářet nové reprezentace (kreativní schopnost) a jednak jako podmínku možnosti racionální kognice ve smyslu kontrafaktuálního uvažování a plánování budoucnosti.

Buckner připomíná, že v rámci Humovy empiristické teorie obrazivosti platí, že mysl je schopná vytvářet nové ideje z materiálu v podobě impresí, které byly dříve přijaty smysly. Tyto imprese jsou zkopírovány pomocí paměti a přetvořeny do podoby idejí (tzv. *copy principle*). Mysl ideje následně rozkládá, skládá a rekombinuje do podoby nových percepcí.[5] Například percepce jednorožce podle Huma vzniká kombinací ideje koně a ideje rohu.

Buckner v kapitole „Humovo pojetí imaginace" hájí koncepci skotského filosofa vůči kritice J. Fodora[6] a poté jej vztahuje ke Generativním adversariálním sítím (GAN), variačním auto-enkoderům (VAE) nebo transfomerorovým architekturám apod. Tyto systémy dle Bucknera modelují určité aspekty Humova pojetí imaginace. Jsou schopné vytvářet nové reprezentace díky rekombinaci dříve přijatého materiálu. Vykazují tak charakter kompozicionality a produktivity (byť nikoliv neomezené).

Zdá se nicméně, že tak nečiní stejným způsobem jako lidé. I přes tuto potenciální námitku chce Buckner alespoň poukázat na skutečnost, že neuronové sítě dokáží funkce této mohutnosti implementovat, byť je nejisté, nakolik je tato implementace biologicky plausibilní.[7] V budoucnu by do systémů strojového učení modelujících racionalitu měly být dle Bucknera kromě kreativity implementovány též (výše zmiňované) schopnosti kontrafaktuálního uvažování a racionálního plánování budoucnosti.[8]

5 Pozn.: Jedná se o vlastnosti kompozicionality a produktivity. Viz níže.

6 Viz Fodor, J. A., *The Modularity of Mind: An Essay on Faculty Psychology*, c.d.

7 Podstata námitky spočívá ve skutečnosti, že v současnosti neexistuje dostatečné silná empirická opora teze hájící případnou podobnost mechanismů imaginace u lidí a výše uvedených systémů.

8 Buckner v rámci své interpretace nezohledňuje možné rozdělení imaginace na schopnost fantazie a schopnost, která svou funkcí zhruba odpovídá kantovské transcendentální obrazotvornosti. Toto rozlišení lze vysledovat v Humově textu a je tematizováno u některých interpretů: Viz např. Allison, H. E., *Custom and Reason in Hume: A Kantian Reading of the First Book of the Treatise*. New York, Oxford University Press 2008. Viz též Hume, D., *Pojednání o lidské přirozenosti. Kniha I: Rozum*. Přel. H. Janoušek. Praha, Togga 2015. Zahrnutí druhé

6. V šesté kapitole Buckner zkoumá pozornost. Tuto mohutnost přitom nahlíží prizmatem teorie W. Jamese – resp. její interpretace u J. Prinze – a následně jeho pojetí usouvztažňuje s mechanismy pozornosti (*self-attention*) v transformerových architekturách (jako jsou GPT či BERT).

Podle Jamese je pozornost aktivním procesem, který dokáže dynamickým způsobem filtrovat informace a zaměřovat se pouze na to, co je důležité. Mechanismus *self-attention* umožňuje analogicky např. transformerovým architekturám GPT přiřazovat různou váhu různým slovům v dané sekvenci slov (resp. přesněji tokenům, ze kterých slova sestávají). Díky tomu dokáží tyto architektury zlepšit svou schopnost „chápat" kontext.

Implementace pozornosti by dle Bucknera mohla napomoci při budování racionálních systémů i v tom, že tato mohutnost umožní řízení kognitivních procesů a efektivní distribuci zdrojů mezi ostatní moduly DoGMA (tzv. problém kontroly).

7. V poslední kapitole pak Buckner zkoumá ještě spektrum schopností spadajících pod sociální kognici v pojetí Adama Smithe a zejména Sofie de Grouchy.

Celkově vzato Buckner v publikaci ukazuje, že umírněný empirismus a psychologie mohutností jsou schopné potenciálně vygenerovat zajímavý a plausibilní teoretický rámec, který umožňuje koherentní popis racionality a kognice a který lze zároveň využít jako „obecnou mapu" pro designování komplexních inteligentních systémů založených na hlubokém strojovém učení.

Namísto snahy o modelování doménově-specifických (tj. úzce specializovaných) kognitivních struktur a mechanismů Buckner navrhuje strategii, která je založená na implementaci mentálních mohutností. Ty jsou přitom v rámci Bucknerova přístupu chápány jako doménově-obecné mechanismy či moduly, které vzájemně interagují a kognici spoluutvářejí aktivním způsobem.

Buckner oceňuje úspěchy existujících architektur hlubokého strojového učení, ale zároveň je kritizuje. Namítá, že ačkoli jsou při vykonávání dílčích úloh velmi efektivní a užitečné, implementují jednotlivé kognitivní funkce izolovaně, a racionalitu proto nemodelují dostatečně komplexním způsobem. Různé aspekty kognice je dle něj třeba z této izolace vytrhnout a shrnout je pod jednou střechou.

z výše uvedených funkcí imaginace by přitom Bucknerovi mohlo v jeho projektu DoGMA pomoci. Humovo pojetí tohoto aspektu obrazivosti totiž (v dané interpretaci) implikuje, že obrazivá mohutnost aktivně strukturuje a dotváří naší percepci prakticky v reálném čase. Bucknerovsko-lockovské pojetí percepce by tedy mohlo být o tuto kompetenci potenciálně rozšířeno. Rozsáhlejší expozici by si zasloužilo také Humovo pojetí víry ve skutečnost (belief) doprovázející určité percepce, které je pro Bucknerův projekt relevantní, avšak v jeho textu je načrtnuto pouze v hrubých obrysech. Hume mimo jiné detailně zkoumá též problém kauzální inference, který je – jak se zdá – pro řadu systémů DNN stále palčivý. V Humově řešení tohoto problému, které se zakládá na síle habituálního spojení, by Buckner rovněž mohl najít vítanou inspiraci.

To by mělo přinést radikální posun pro další fáze rozvoje umělé racionality vyšších úrovní. Právě v tomto ohledu může DoGMa – jakožto obecný teoretický rámec – dle Bucknera pomoci.

Možnou slabinou knihy je empirická opora tezí o analogičnosti procesů u lidí a DNN. Ta je nejsilnější ve třetí kapitole a poté se s každou další kapitolou snižuje.[9] Buckner rovněž explicitně netematizuje existenci řady dalších kognitivních architektur, přičemž některé z nich usilují o podobně komplexní přístup k modelování kognice jako on.[10] Třetí potenciální námitka pak vychází ze skutečnosti, že Bucknerovo pojetí architektury DoGMA je velmi obecné. Jelikož tato architektura jako celek není spojena s žádnou komputační implementací, jedná se – alespoň prozatím – o obecný návrh[11] existující pouze v rovině teorie. Tento návrh může nicméně ve výsledku ukázat, jaké další kroky bychom měli při vývoji komplexních systémů umělé racionality podniknout.

Publikace bude přínosná zejména pro filosofy zabývající se kognicí, kognitivní vědce a výzkumníky v oblasti AI.

Jednoznačně ji doporučuji k přečtení.

9 Srov. Buckner, J., *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artifical Intelligence*, c.d., s. 203.

10 Viz Kotseruba, I. – Tsotsos, J. K., 40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications. *The Artificial Intelligence Review*, 53, 2020, No. 1. Viz též Ye, P. – Wang, T. – Wang, F.-Y., A Survey of Cognitive Architectures in the Past 20 Years. *IEEE Trans Cybernetics*, 48, 2018, No. 12. Dostupné na: http://doi:10.1109/TCYB.2018.2857704; [cit. 24. 10. 2025].

11 Byť je podpořený sérií zajímavých (avšak spíše dílčích) analogií mezi empiristickými teoriemi mohutností a systémy DNN.