TEORIE VĚDY / THEORY OF SCIENCE / 2024
https://doi.org/10.46938/tv.2024.631

////// recenze / review //

PROBLÉM SLADĚNÍ

Christian, Brian. *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* London: Atlantic Books, 2021. 476 stran.

TEREZA MATĚJKOVÁ
Katedra filozofie
Západočeská univerzita v Plzni
Sedláčkova 19, 306 14 Plzeň
email / matejkot@ff.zcu.cz

V knize *The Alignment Problem,* s podtitulem Machine Learning and Human Values, nám Brian Christian[1] nabízí fascinující a hluboký pohled do oblasti strojového učení a jejího vývoje. Tato monografie, jak naznačuje název, se zaměřuje na v současné době klíčový problém, kterým je problém sladění (*alignment problem*). Ten se týká způsobu, jakým zajistit, aby systémy umělé inteligence vykazovaly chování v souladu s lidskými hodnotami, etikou a cíli. Autor nezkoumá abstraktní hodnoty lidstva jako celku, ale zabývá se zejména cíli a přáními jednotlivců, kteří by s těmito systémemy pracovali. Základním cílem vyřešení problému sladění je dosažení soudržnosti mezi stanoveným cílem systému a výsledným chováním tohoto systému. Brian Christian si klade ambiciózní úkol: prozkoumat a analyzovat problém sladění a zvýšit tak povědomí o problematice a komplexnosti úsilí uvést v soulad systémy strojového učení a umělé inteligence s lidskými hodnotami.

Tato kniha, ačkoli poprvé publikována v roce 2020, je dle mého názoru stále jednou z nejpovedenějších a nejpoučnějších knih o historii vývoje v oblasti umělé inteligence.[2] Nabízí souhrn klíčových momentů v historii strojového učení, které vedly k současnému stavu poznání. Kousek po kousku

[1] Brian Christian (*1984) vystudoval informatiku a filosofii na Brown University, poezii na University of Washington a psychologii a komputační neurovědu na University of Oxford. Ve své výzkumu a díle se věnuje důsledkům, které mají počítačové vědy pro člověka. Jeho dalšími díly jsou knihy *The Most Human Human* z roku 2011 a *Algorithms to Live By* z roku 2016.
[2] Dalšími knihami, které bych do této kategorie zařadila, jsou např. Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* z roku 2019; Melanie Mitchell, *Artificial Intelligence: A Guide for Thinking Humans* z roku 2019 a Max Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence,* z roku 2017.

představuje kroky, díky nimž dnes víme, na co si dát při právě probíhajícím rozmachu umělé inteligence pozor a případně jaké otázky klást (kdo/co je reprezentován v trénovacím datasetu, za co je systém odměňován, dokáže systém vysvětlit kroky, které vedly k výslednému rozhodnutí atd). Ovšem tyto klíčové momenty autor nepředkládá jen jako výčet důležitých informací a poznatků, ale propojuje je prostřednictvím snahy o sladění lidských a „strojových" hodnot a cílů.

Autor předesílá, že kniha je výsledkem stovky formálních a několika stovek neformálních rozhovorů, které vedl během čtyř let s výzkumníky a výzkumnicemi z oblasti počítačové vědy. Avšak výsledný text neposkytuje pouze vhled do počítačové vědy. V předložené monografii se technologické aspekty prolínají s filosofií, psychologií, neurovědou a také historií těchto oborů. To vše pak formuje vyprávění „o strojovém učení a lidských hodnotách: o systémech, které se učí z dat bez explicitního programování, a o tom, jak vlastně a co přesně se je snažíme naučit" (s. 11). Tato publikace v sobě tedy snoubí vyprávění příběhu strojového učení, který je předáván prostřednictvím osobních příběhů a rozhovorů, se srozumitelným výkladem technických aspektů umělé inteligence. Autor přesně nestanovuje, pro koho knihu zamýšlel, ale dle mého názoru je relevantní pro každého, kdo sice má už minimálně laické počáteční znalosti tohoto oboru, ale chce svůj přehled prohloubit o souvislosti, které lze nalézt mezi disciplínami kognitivní vědy. Velká část této knihy je ukazováním různých problémů, které vyvstaly při rozvoji strojového učení a umělé inteligence, a jejich řešení, které je často inspirované psychologií či neurovědou, což odráží i podrobnější obsahová struktura knihy. Christian postupuje chronologicky s vývojem počítačových věd a vždy v rámci daného tématu předloží konkrétní problém, následně ukáže, jak byl tento problém vyřešen, a hned navazuje předložením nově vzniknuvšího problému.

Kniha je systematicky strukturována do tří částí. V první z nich, pojmenované *Prophecy* (s. 15–117), se autor věnuje historii výzkumu systémů strojového učení a odhaluje některé jejich nezamýšlené a nepředvídané projevy, s nimiž se odborníci a odbornice při vývoji setkali. Christian se věnuje především prediktivní schopnosti systémů strojového učení a umělé inteligence, tomu, jak tyto technologie „předpovídají" budoucí chování na základě předložených dat a také selekčnímu zkreslení, které ilustruje na konkrétních příkladech. Čteme tak o tom, jak nedostatečná diverzita dat může vést k předpojatým výsledkům a stereotypům, a to jak v obrazových,

tak i v jazykových modelech[3] (s. 31). Na základě toho varuje před nerozvážným používáním jazykových modelů, které by mohly udržovat či umocňovat předsudky ve společnosti. Podle autora by při správném a optimálním používání naopak tyto modely mohly přispět k překonání zakořeněných předsudků (s. 38–40).

Od problematiky reprezentace se Christian dostává k tématu spravedlnosti (s. 51–81) v kontextu predikčních modelů používaných v trestním soudnictví, zejména při hodnocení rizika recidivy. Upozorňuje na důležitost správné interpretace toho, co modely skutečně předpovídají, a poukazuje na nebezpečí, že modely mohou být trénovány na datech, která neodrážejí skutečnou realitu, ale spíše zkreslené lidské vnímání nebo nespravedlivé vzorce vymáhání práva. Na základě prostudovaných výzkumů v kombinaci s osobními rozhovory zdůrazňuje nutnost kritického přístupu k využívání predikčních modelů v trestním soudnictví. Prostřednictvím slov Moritze Hardta a odkazem na Ernesta Burgesse uzavírá, že změna potřebná ke snížení kriminality a zvýšení efektivnosti soudního a vězeňského systému musí vycházet ze strukturální změny ve společnosti.

Christian se při poukazování na nedostatky reprezentace a spravedlnosti v systémech strojového učení drží v mezích potřebných pro vysvětlení technického problému, a ačkoli na skutečný společenský problém odkazuje, dělá to hlavně prostřednictvím příběhů a hlasů jiných odborníků a odbornic. Nenabízí tedy hlubší pohled na příčiny těchto problémů, což je pochopitelné, jelikož sám autor není odborníkem na sociologická nebo právní témata, a proto se přirozeně soustředí na technické aspekty.

Dále navazuje požadavkem transparentnosti a vysvětlitelnosti modelů založených na strojovém učení. Autor představuje hned několik možných přístupů, jak docílit toho, aby modely byly lépe vysvětlitelné, a tudíž aby bylo možné více důvěřovat jejich výstupům. Jedním ze způsobů, které Christian popisuje, je dekonvoluce,[4] kterou vyvinuli Matthew Zeiler a Rob Fergus. Sami v současné době můžeme pozorovat, že si stále větší pozornost získává oblast tzv. vysvětlitelné umělé inteligence (XAI), tj. oblast umělé

[3] Jedním z příkladů může být databáze obličejů Labeled Faces in the Wild, kterou v roce 2014 analyzovali Hu Han a Anil Jain se závěrem, že je tvořena ze 77 % mužskými obličeji a z 83 % obsahuje bělošské tváře. Viz Hu Han and Anil K. Jain, *Age, Gender and Race Estimation from Unconstrained Face Images*, MSU Technical Report (MSU-CSE-14-50), East Lansing: Michigan State University, July 2014.

[4] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. „Deconvolutional Networks," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR) (San Francisco, CA: IEEE, 2010): 2528–35.

inteligence zaměřená na vývoj systémů, které, kromě provedení úkolů, také poskytují srozumitelné vysvětlení svých rozhodnutí a akcí. Cílem XAI je transparentnost modelů umělé inteligence, aby uživatelé či vývojáři mohli lépe porozumět tomu, jak tyto systémy dospívají ke svým závěrům.

Druhá část knihy nese název *Agency* (s. 119–210) z toho důvodu, že se v ní autor soustředí na otázku, jak systémy umělé inteligence jednají a přijímají rozhodnutí. Je zde popsána historie výzkumu učení podmiňováním, učení posilováním a vnitřní motivace, které se prolínají s počítačovou vědou a vývojem systémů umělé inteligence. Christian v kontextu Atari her a deskové hry Go prezentuje, jakým způsobem se strojové učení inspirovalo poznatky psychologie s cílem vytvářet takové systémy, které budou vykonávat činnosti a rozhodovat se autonomně. Je třeba ocenit, v jaké detailnosti jsou představeny počátky výzkumu Edwarda Thorndikea a jeho známého „zákonu účinku" (*law of effect*). Tento zákon zjednodušeně říká, že chování, které je následováno uspokojivými důsledky má tendenci se opakovat, kdežto množství chování, které je následováno nepříjemnými důsledky, se snižuje. Tento princip je pak klíčovým konceptem ve zpětnovazebném učení (*reinforcement learning*), který umožňuje strojům nebo agentům[5] učit se z interakce s prostředím pomocí zpětné vazby ve formě odměn a trestů. Způsob, jakým Christian přechází mezi strojovým učením, pokusy se zvířaty a dětmi a fungováním mozku, dodává textu, kromě zajímavosti a čtivosti také soudržnost a přidanou informativní hodnotu, protože autor poukazuje na vzájemné souvislosti získávaných poznatků. Čteme tak například o pokrocích ve výzkumu Richarda Suttona a Arthura Samuela, které vedlo Geralda Tesauro k vytvoření průlomového programu TD-Gammon[6] (s. 141), jež byl jedním z prvních programů, který úspěšně využil metodu zpětnovazebného učení v kombinaci s neuronovými sítěmi.

Christian se následně zabývá typem učení zvané tvarování (*shaping*) a jeho využitím pro překonání problému řídkosti (*problem of sparsity*), tj. když je odměna nastavená tak, že k ní dojde až v cíli. K tomuto cíli se nakonec stroj nebo agent sice dostane nějakým náhodným pohybem, ovšem může to trvat extrémně dlouho. Řešením se tedy stalo odměňování jednotlivých malých dílčích kroků, které postupně vedou ke kýženému výsledku.

[5] V kontextu strojového učení a umělé inteligence je agent označení pro počítačový program nebo systém, který je navržen tak, aby vnímal své prostředí, rozhodoval se a prováděl akce k dosažení určitého cíle nebo souboru cílů. Agent pracuje autonomně, což znamená, že není přímo řízen lidským operátorem.

[6] Gerald Tesauro, „TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play," *Neural Computation* 6, no. 2 (1994): 215–19.

Opět na konkrétních případech ukazuje, jak významnými jsou při učení, ať už strojovém nebo lidském, správně navržený plán učení a vhodně zvolené pobídky. Nahlíží na ně z několika perspektiv a prezentuje, co vše je potřeba si uvědomit při tvorbě a trénování systému. Za nejdůležitější považuje nutnost rozlišovat to, co chceme, aby agent dělal, a to, za co ho průběžně odměňujeme.

Oblast strojového učení se však neinspiruje jen způsoby vnější motivace, ale také mechanismy vnitřní motivace, což znamená využívání mechanismů, jako je novost a překvapení, které podněcují algoritmy k objevování a učení z nových, neznámých nebo nečekaných aspektů dat. Tento přístup reflektuje principy vnitřní motivace v psychologii, kde jsou jednotlivci poháněni osobním uspokojením či zájmem o novost. Christian problematiku novosti a překvapení vysvětluje na chování dětí (s. 182–210), přičemž se odkazuje na významné výzkumy druhé poloviny 20. století. Při absenci silných pobídek se dítě nechová náhodně, ale přitahují jej spíše nové věci, které ho něco překvapivého naučí. Obor umělé inteligence se tímto inspiroval například při vývoji agentů, kteří jsou odměňováni za vytváření překvapivých předpovědí. Tuto část knihy uzavírá diskuze o výzvách při navrhování systémů s vnitřní motivací, přičemž autor tímto způsobem naznačuje, že pominutí vnějších odměn může být klíčové pro vytvoření skutečně obecné umělé inteligence. Právě tímto směrem by se dle Christiana měl ubírat další výzkum umělé inteligence.

V průběhu celé této části správně upozorňuje na rozdílnosti skutečného světa a modelového prostředí, v němž se strojové učení odehrává. Velký důraz klade na fakt, že umělý agent nemůže udělat fatální chybu, tj. takovou, která jej zničí a nedovolí mu se obnovit, a také na fakt, že v reálném světě proměňujeme naše cíle v reakci na prostředí a ostatní agenty a že dokážeme svým chováním svět formovat. Christian uvádí, že většina systémů strojového učení předpokládá, že sami svět nijak neovlivňují, tudíž do svého modelu nezahrnují sami sebe.

Poslední třetina monografie nazvaná *Normativity* (s. 211–310) je zaměřena na trénování umělé inteligence prostřednictvím imitace lidského chování. Je zde přehledně představeno učení imitací a jeho tři hlavní výhody: efektivita, možnost naučit se obtížně popsatelné jednání a bezpečnost, protože dochází k napodobování již ověřených vzorců chování. Ovšem i tento typ učení má některé problematické aspekty, které je třeba zohlednit. Například pokud se agent učí od experta, nenaučí se, jak napravit chyby, které by případně udělal. Proto je důležitá interakce mezi expertem a učícím se agentem v problematických situacích. Christian tedy zohledňuje potenciální problematické aspekty imitace a načrtává možnosti jejich řešení.

Zůstávají tak otevřeny jen ty otázky, na něž současný výzkum zatím nedokázal poskytnout odpovědi. Nutno dodat, že tyto odpovědi neposkytuje ani Christian, spíše jen opatrně naznačuje směr, kterým je dle něj dobré se vydat při řešení problému sladění.

Důležitým stupněm v cestě za vyřešením problému sladění je podle něj také schopnost odvozování záměrů, která je pro lidské poznání zásadní. Autor navazuje představením inverzního učení posilováním (*inverse reinforcement learning*) a kooperativního inverzního učení posilováním (*cooperative IRL*), které považuje za klíčový přístup, jenž by mohl vést k vyřešení problému sladění. Znamená to, že by se stroje mohly učit lidským hodnotám pozorováním našeho chování a nespoléhat pouze na explicitní naprogramování. Vedou ho k tomu úspěchy těchto přístupů, jako příklad můžeme uvést kinesthetic teaching.

V poslední části knihy (s. 277–310) se autor zabývá problematikou nejistoty v systémech strojového učení, která vychází z jejich omezené znalosti světa. Řešení tohoto problému představuje používání pravděpodobnostních neuronových sítí. Pro Christiana je zásadní zachovat si možnost intervenovat a systém korigovat. Upozorňuje však na okruh problémů, které se váží k systému, jenž nejistotu vykazuje. Jedním z problémů je například ztráta nejistoty zapříčiněná korekcí či potvrzením výstupu, která může vést k neposlušnosti systému. Za řešení této situace je považován tzv. IRD (*inverse reward design*), v jehož rámci jsou lidské pokyny považovány za nedokonalé obrazy jejich skutečných přání. Cílem je zajistit, aby systémy umělé inteligence nebraly instrukce doslovně, ale používaly je jako informativní signály při zachování nejistoty, která usnadňuje kontrolu a dodržování pokynů.

Když to tedy znovu shrneme, monografie Briana Christiana nás seznamuje se sílou a možnostmi modelů strojového učení, s jejich nedostatky, možnými problematickými aspekty a způsoby, jakými dochází k postupnému sladění s našimi zájmy, cíli a přáními. Tento velký příběh pokroku strojového učení a lidských hodnot se snaží zachytit na docela malém prostoru (334 stran textu), což znamená zhuštění informací do takové míry, že může být obtížné udržovat v paměti veškeré předkládané poznatky a souvislosti. Ovšem to, že je výklad opřen o vysvětlování konceptů na konkrétních příkladech, experimentech, studiích a osobních rozhovorech, jej činí stravitelnějším. Díky tomu je úspěšný ve vysvětlování technických konceptů pochopitelnou formou jak pro humanitně, tak pro technicky zaměřené čtenářstvo. Mimo jiné Christian také ukazuje, jak si výzkumníci a výzkumnice v oblasti umělé inteligence postupně více a více uvědomují, že jsou výsledky jejich výzkumů ovlivněny společenskými hodnotami a že

tyto výsledky také dále ovlivňují tyto proměňující se hodnoty, a apeluje na bezpečné užívání těchto systémů, protože problematické nemusí být jen algoritmy, ale především lidé. Svou interdisciplinární povahou, postupným rozvíjením témat, vysvětlováním kontextu a přístupným jazykem se kniha vyhýbá riziku, že by byla odborníky vnímána jako povrchní a laiky jako příliš složitá. Autor dokázal osvětlit souvislosti několika různých oborů zabývajících se učením a upozornil na filosofické a etické problémy spojené s vývojem umělé inteligence.

Vzhledem k tomu, že se problematika sladění systémů strojového učení a umělé inteligence s lidskými zájmy a hodnotami stává stále naléhavější, je nyní důležitější než kdy dříve upozorňovat na tento problém a znát mezery a omezení, které tyto systémy mohou mít. Pozornost k problému sladění přitáhlo například nedávné „rozpuštění" tzv. superalignment týmu v OpenAI, který se měl věnovat právě těmto otázkám. Vyvolalo to v tomto směru určité obavy z dodržování etických norem.[7] Je proto klíčové, aby technologie odpovídala etickým normám a očekáváním společnosti a aby nedocházelo k nepředvídatelným a potenciálně škodlivým důsledkům.

Je však třeba poznamenat, že podobně jako u jiných textů o umělé inteligenci i tento se v některých ohledech stal méně aktuálním, což je přirozeným důsledkem závratné rychlosti vývoje a výzkumu v tomto oboru. Ačkoliv se kniha věnuje velmi aktuálnímu tématu, nezahrnuje například nedávný vzestup generativní umělé inteligence, který zaznamenáváme především v posledních dvou letech s nástupem ChatGPT. I přesto však považuji tuto monografii za důležitý příspěvek v této oblasti, a to právě díky popsané historii vývoje a poukázání na souvislosti mezi poznatky různých oborů.

Bibliografie:

Christian, Brian. *The Most Human Human*. New York: Doubleday, 2011.

Christian, Brian. *Algorithms to Live By*. New York: Henry Holt, 2016.

Han, Hu, and Anil K. Jain. *Age, Gender and Race Estimation from Unconstrained Face Images*. MSU Technical Report (MSU-CSE-14-5), East Lansing: Michigan State University, July 2014.

Knight, Will. „OpenAI's Long-Term AI Risk Team Has Disbanded." *Wired*, May 17, 2024. https://www.wired.com/story/openai-superalignment-team-disbanded/.

[7] Will Knight, „OpenAI's Long-Term AI Risk Team Has Disbanded," *Wired*, May 17, 2024.

Mitchell, Melanie. *Artificial Intelligence: A Guide for Thinking Humans*. London: Macmillan, 2019.

Russell, Stuart J. *Human Compatible: AI and the Problem of Control*. New York: Viking, 2019.

Tegmark, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf, 2017.

Tesauro, Gerald. „TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play." *Neural Computation* 6, no. 2 (1994): 215–19.

Zeiler, Matthew D., Dilip Krishnan, Graham W. Taylor, and Rob Fergus. „Deconvolutional Networks." In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR), 2528–35. San Francisco, CA: IEEE, 2010.