OPRAVDU SE MÁME BÁT UMĚLÉ INTELIGENCE?

David ČERNÝ, Jiří WIEDERMANN

Abstrakt
V posledních měsících je umělá inteligence, přesněji některé systémy vykazující inteligentní chování (např. ChatGPT-3.5), na stránkách všech novin. Udivuje svými mimořádnými schopnostmi, ale vzbuzuje i velké obavy. Nedávno se dokonce objevila otevřená výzva několika odborníků a známých osobností, která volá po zastavení vývoje AI na půl roku. Svou polívčičku si přihřívají i známí technopesimisté, kteří považují AI za poslední lidský vynález, který má jasný potenciál přinést lidstvu zkázu. Ve svém příspěvku nejdříve představím některé z těchto technopesimistických postojů, ukáži, jaký vztah spatřují mezi obecnou umělou inteligencí, superinteligencí, singularitou a možnou zkázou lidstva. Následně se pokusím ukázat, že tyto pesimistické předpovědi nejsou příliš opodstatněné. Hlavní problém není umělá inteligence jako taková, ale naše pasivita a neschopnost (či neochota) využívat ji korektním a rozumným způsobem. Nakonec pohovořím o nutnosti kultivovat algoritmickou gramotnost na všech úrovních vzdělávacího procesu.

EXISTENTIAL RISKS ASSOCIATED WITH ARTIFICIAL INTELLIGENCE

Abstract
In this paper we focus on the existential risks associated with artificial intelligence. We try to show that these risks may be greatly exaggerated. In fact, the path from the current state of the AI field to general AI and superintelligence is not as well understood as some authors assume. We are moving into the realm of the unknown, of many "maybes" and uncertainties. Moreover, it is a mistake to pass off these unknowns as knowns, as an unveiled map of the whole territory. We also try to point out that focusing on the existential risks of AI (the territory of uncertainties and unknowns) is not an appropriate strategy because it easily distracts from the issues at hand. AI is ubiquitous and has a profound impact on our understanding of what it means to be human and how we relate to the world and to other humans. We can look at AI as a challenge, as a distinct possibility for cognitive enhancement of our minds (expanded mind), but we need to use the tools of AI in a rational, wise and morally right manner. The greatest risk today is not the AI of the distant future, but our passivity and unwillingness to take our destiny into our own hands.

Moderní systémy vykazující umělou inteligenci, nebo pro jednoduchost umělá inteligence (AI) plní titulky novin a jen málokomu mohlo ujít, že se děje něco velkého. Jako by v posledních měsících došlo k obrovskému skoku. Hluboké neuronové sítě trénované na neuvěřitelně velkých sadách dat, které stojí v pozadí pozoruhodných pokroků, dnes dokáží skládat hudbu, generovat mimořádně kvalitní obrázky, vést s námi konverzaci v našem vlastním jazyce, dokonce umí zkontrolovat práci programátorů a navrhnout zlepšení jejich kódů. A hned se samozřejmě začaly objevovat aplikace, které těchto výdobytků moderní techniky využívají. Mohou například sledovat naše mailové

konverzace a automaticky připravovat odpovědi, poskytovat právní rady, psát smlouvy či fungovat jako naši osobní asistenti.

A alarmisté bijí na poplach.[4] V pokrocích na poli umělé inteligence nevidí nástroj, který nám může úžasně zlepšit život, ale především hrozbu, která nám všem jednou zakroutí krkem. Britský filosof Toby Orb se dokonce pokusil tuto hrozbu vyjádřit exaktně a došel ke strašidelnému poměru 1 : 10.[5] Právě taková je podle něj šance, že se AI vydá na cestu svých předchůdců v mnoha dílech science fiction, jako byl neblaze proslulý palubní počítač HAL 9000, Skynet a jeho terminátoři, Ultron známý z marvelovských komiksů či třeba Cyloni vedoucí vyhlazovací válku s posledními zbytky lidstva. Podle mnohých jsme stvořili Frankensteinovo monstrum: úžasně silné a mocné, ale mající své vlastní plány, jejichž realizaci mohou lidé překážet. Vzpoura vůči stvořiteli je koneckonců jedním z hlavních námětů lidských dějin. Zatímco ale křesťanského, židovského či muslimského boha naše vzpoura nijak neohrožuje, Frankensteinovy systémy umělé inteligence nás mohou přivést na pokraj vyhubení, nebo se nás zbavit jednou provždy.

Jak by se jim to ale mělo povést a hlavně, proč by proboha něco takového umělá inteligence dělala? Technopesimisté, jak jim můžeme říkat, mají celou řadu odpovědí. Nebude to snadná cesta, ale je podle nich jasně narýsovaná a lidstvo se po ní již řítí nezvladatelnou rychlostí. Nejdříve stvoříme obecnou umělou inteligenci (GAI)[6], která

[4] Slovo „alarmisté“ zde nemíníme v nějakém negativním smyslu. Uvědomujeme si také, že se jedná o velmi širokou kategorii lidí, mezi nimiž by bylo vhodné dále rozlišovat. Zahrnuje intelektuály z různých oborů, ředitele velkých společností, ale i filosofy a odborníky na umělou inteligenci. Za alarmisty zde označujeme ty, kteří se domnívají, že umělá inteligence představuje významné, možná i existenční, riziko, a upozorňují na to ve svých publikacích a veřejných vystoupeních. Mezi hlavní postavy alarmistů lze zařadit Nicka Bostroma (BOSTROM, Nick. *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press, 2016); S. Russella (RUSSELL, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. London: Penguin, 2020); Maxe Tegmarka (TEGMARK, Max. *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Allen Lane, an imprint of Penguin Books, 2017) a Tobyho Orda (ORD, Toby. *The Precipice: Existential Risk and the Future of Humanity*. New York: Hachette Books, 2020).

[5] ORD, Toby. *The Precipice: Existential Risk and the Future of Humanity,* cit. Diskuse o AI jako existenčním riziku se nachází na s. 138−152, vyjádření šance, že se vyjmenovaná rizika aktualizují, se nachází na s. 167.

[6] Je třeba říci, že názory odborníků se velmi rozcházejí ohledně předpovědi, zda je GAI vůbec možná (většina si ale myslí, že ano), jakou bude mít podobu („v počítači“ – ztělesněná (robotická)) a kdy se jí dočkáme. Alarmisté často vyjadřují názor, že GAI je v podstatě „za rohem“, zatímco jiní autoři zdůrazňují, že existuje celá řada důležitých kognitivních dovedností (abdukce, schopnost učení se v reálném, dynamickém a nesmírně komplexním prostředí fyzického světa), v nichž současná AI zatím zásadně selhává. Například J. Landgrebe a B. Smith se domnívají, že AI nikdy obecné inteligence nedosáhne (LANDGREBE, Jobst a Barry SMITH. *Why machines will never rule the world: artificial intelligence without fear*. New York, 2023), před humbukem ohledně možností AI varuje také odborník na hluboké učení F. Chollet (CHOLLET, François. *Deep learning with Python*. Second edition. Shelter Island: Manning, 2021), na důležitost některých forem lidské inteligence, jimiž stroje v lidské míře ani zdaleka nedisponují, upozorňují např. G. Marcus a E. Davis (MARCUS, Gary F. a Ernest DAVIS. *Rebooting AI: building artificial intelligence we can trust*. New York: Pantheon Books, 2019) či Erik J. Larson, který se soustředí zvláště na abduktivní inference (LARSON, Erik J. *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*. Cambridge, Mass.: Belknap Press, 2021). Je třeba si uvědomit, že spor o tom, zda se blížíme k GAI není ve své podstatě sporem čistě vědeckým, ale do značné míry filosofickým. A rozdíly v názorech mezi odborníky vycházejí z jejich odlišných filosofických představ o tom, co je to obecná inteligence a zda existují nějaké znaky (a jaké), jejich splňováním se k GAI blížíme či jsme dokonce již velmi blízko. K tomu srov. SUMMERFIELD, Christopher. *Natural General Intelligence: How understanding the brain can help is build AI*. New York: Oxford University Press, 2023.

bude zvládat všechno, co zvládá inteligence lidská. Možná jinak, zřejmě i rychleji, ale jinak se od ní ve svých schopnostech lišit nebude. A dostane od nás jeden velmi specifický úkol (nebo si ho tajně uloží sama)[7]: zaměř se na svou vlastní architekturu, začni ji vylepšovat a vytvářej stále inteligentnější stroje. Asi jako slavná Hlubina myšlení, která sice objevila odpověď na základní otázku života, vesmíru a vůbec, ale nevěděla, jak ta otázka vlastně zní. Proto navrhla nejmocnější počítač ve známém vesmíru, jímž zvláštní shodou okolností byla planeta Země, jejíž na uhlíku založení dvounozí vládci o tom neměli ani tušení. Nakonec je ale nezlikvidovala umělá inteligence, nýbrž potměšilá a krutá rasa Vogonů, která se vyžívá v mučení ostatních předčítáním vlastní poezie.

GAI nakonec, možná dokonce velmi rychle, uspěje a dojde ke vzniku umělé superinteligence (SAI) a vývoj doputuje do singularity.[8] Tu můžeme chápat různě. Jako bod, za nímž bude následovat inteligenční exploze: povstávání nových a stále (super)inteligentnějších strojů. Nebo jako okamžik, za kterým se vývoj techniky stane pro naše omezené lidské mozky zcela nepochopitelným a netransparentním a ztratíme tak nad ním veškerou kontrolu. Anebo, budeme-li se držet základního vyprávění technopesimistů, jako hranici, za níž je už náš osud zpečetěn.

AI je často představována jako existenční riziko, tedy takové riziko, které – pokud by se naplnilo – by buď úplně vyhubilo všechny lidské bytosti, nebo by „alespoň" redukovalo lidskou populaci a civilizaci na úroveň, v níž není prostoru pro nic specificky lidského, jako je kultura, právo, morálka, kultivace poznání apod.[9] Naše planeta byla v kultovní sci-fi Douglase Adamse *Stopařův průvodce po galaxii* nejdokonalejším systémem umělé inteligence; přesto lidstvo nevyhubila. Dokonce ho potřebovala, stejně jako lidstvo potřebovalo ji. Vzájemně koexistovali a nebýt nutnosti výstavby

[7] Proč by si AI měla uložit úkol (tajně) vylepšovat svou kognitivní strukturu a vytvářet své inteligentnější kopie? Jednou z možností je to, že si to určí jako instrumentální cíl. Je-li F nějaký cíl jednání, potom struktura S cílů C_1,…, C_n je strukturou instrumentálních cílů, pokud C_1,…, C_n přispívají (jsou podmínkou) realizace F. Řekněme např., že lidé zadají stroji cíl F = najdi lék na rakovinu. Stroj začne F řešit a uvědomí si, že mezi instrumentální cíl patří i vylepšení vlastních schopností; začne je tedy vylepšovat. K tomu srov. BOSTROM, Nick. *Superintelligence: paths, dangers, strategies*, cit., kap. The superintelligent will. Bostrom se dokonce domnívá, že některé cíle jsou ze své povahy takové, že se jich pravděpodobně pokusí dosáhnout takřka každá inteligence (toto tvrzení vyjadřuje tezi o instrumentální konvergenci, s. 131−132).

[8] Technické singularitě se věnují dvě důležité kolekce textů: AWRET, Uziel, ed. *The Singularity: Could artificial intelligence really out-think us (and would we want it to)?*. Exeter: Imprint Academic, 2016; EDEN, Amnon H., et al. ed. *Singularity Hypothesis: A Scientific and Philosophical Assessment*. Heidelberg: Springer, 2012. Je třeba říci, že ne všichni autoři vyhlížejí singularitu s obavami. Pro některé je bodem, za kterým nastane neuvěřitelně rychlý (a zrychlující) vědeckotechnický vývoj, který je klíčem k našemu přežití a rozšiřování za hranice Sluneční soustavy. Srov. KURZWEIL, Ray. *The Singularity Is Near: When Humans Transcend Biology*. London: Penguin Books, 2005. Kolem Kurzweilovy interpretace singularity, jejímž jádrem je představa, že lidé musí nakonec splynout se stroji a věda a technika vytvoří ráj na zemi, vznikla komunita, která má charakter nového náboženského hnutí. Srov. GERACI, Robert M. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. New York: Oxford University Press, 2010.

[9] Rizika (možnosti, že nastane něco špatného) lze klasifikovat různě, např. na základě jejich pravděpodobnosti, intenzity a rozsahu. Přidržíme-li se posledních dvou proměnných, potom lze rozlišit rizika: i) podle rozsahu: a) osobní (zasahující jen jednoho člověka), b) lokální, c) globální a d) transgenerační; ii) podle intenzity: a) mírná, b) snesitelná, c) terminální. Existenční rizika jsou rizika, která jsou 1. terminální, 2. transgenerační. Jinými slovy, pokud by nastala nějaká událost klasifikovaná jako existenční riziko, došlo by buď a) k vyhubení lidstva, nebo ii) drastické redukci lidské populace, která by mohla přežívat v podmínkách velmi vzdálených dnešním, bez vědy, techniky, kultury, respektu k lidským právům a hodnotám. Srov. BOSTROM, Nick a Milan M. ĆIRKOVIĆ. Introduction. In: BOSTROM, Nick a Milan M. ĆIRKOVIĆ. *Global Catastrophic Risks*. Oxford: Oxford University Press, 2012, s. 1−29.

intergalaktické dálnice, mohla Země dokončit hledání na základní otázku. Proč by tedy AI měla být vůči nám nepřátelská?

Vtip je v tom, že taková být nemusí. Alespoň si to myslí technopesimisté jako je oxfordský filosof Nick Bostrom, autor bestseleru *Superinteligence*. Úplně stačí, když se zaměří na splnění nějakého úkolu, a tak nějak nezvládne dobře promyslet vhodné prostředky. Když jí třeba poručíme, aby učinila všechny lidi šťastnými, může si pomyslet, že by to dokonale zvládla tím, že nás všechny zdroguje. Mohlo by ji také napadnout – jako některé moderní antinatalisty – že zdrojem utrpení je lidská existence a nejlepším způsobem jeho vymýcení je eliminace trpících. Nebo když ji požádáme, aby vyřešila problém klimatické změny, může dojít k přesvědčení, že nejlepším způsobem je zbavit se všech znečišťovatelů, což jsme shodou okolností zase my lidé. Asi na nás bude také shlížet spatra, jako to ostatně děláme i my, když jde o nějaké „nižší" inteligence, a nebrat naše potřeby a zájmy dostatečně v úvahu. Když člověk staví přehradu, mravenci mají smůlu. Většinou nám nepřijdou natolik důležití, abychom přehradu postavili jinde nebo třeba jejich mraveniště opatrně přemístili jinam.

Příběh o „zlé" AI je tak vlastně povídáním o hloupé AI (je sice superinteligentní a sečtělá, ale nějak jí nedojde, že zdrogovat všechny lidi není moc dobrý nápad) nebo nevšímavé AI (unikne jí, že z etického hlediska možná existuje ohromný rozdíl mezi inteligencí lidskou a mravenčí, ale již mnohem menší (možná skoro žádný) mezi inteligencí lidskou a umělou superinteligencí) nebo popletené AI (bude do zblbnutí vyrábět svorky na papír, dokud nezničí všechny zdroje na Zemi a tím i nás, a navzdory její superinteligenci jí nedojde, jaká je to hloupost).[10]

Příběh je to nepochybně chytlavý; všechna média se ho ostatně chytají jako moucha na med. Je ale věrohodný? A neodvádí náhodou naši pozornost od skutečných problémů?

Současné umělointeligentní systémy jsou nepochybně pozoruhodné a dokážou věci dříve nevídané. Přes to přese všechno jsou to ale pouhé programy, sofistikované, nepochybně, ale stále jsou na hony vzdálené lidské inteligenci. Vezměme si například velký jazykový model společnosti OpenAI, GPT-4. Když s ním komunikujeme, jako by k nám promlouval jiný člověk, jako kdyby rozuměl našim dotazům a snažil se je co nejlépe zodpovědět. Jsou to ale jen masivní umělé neuronové sítě, které dokáží provádět rozsáhlé statistické analýzy nad soubory dat (texty) a umně skládat slova do vět.[11] Ty sítě nerozumí, nevnímají, po ničem netouží, o nic se nesnaží, pouze mechanicky uplatňují pravidla, která se jim pod dohledem člověka podařilo objevit. A jsou samozřejmě omezené svou architekturou a daty, na nichž jsou trénovány. Nemůže se stát, že by se tato síť najednou rozhodla, že začne potají napadat bankovní účty, převádět si finanční prostředky na své účty a financovat tak vlastní armádu strojů. Nebo že by si usmyslela, že začne potají vytvářet své lepší a inteligentnější verze, které bude skrývat v kyberprostoru a jednou z nich povstane nový vládce světa. Nic takového se jednoduše nestane.

Netvrdíme, že se jednoho dne k obecné umělé inteligenci nedopracujeme. Zatím se k ní ale příliš neblížíme a strachovat se o to, jak se bude chovat a zda její další produkty

[10] Srov. např. skeptické hodnocení hrozeb spojených s AI z pera Stevena Pinkera (PINKER, Steven. *Enlightenment Now.: The Case for Reason, Science, Humanism and Progress*. London: Penguin Books, 2018, kap. Existential Threats).

[11] Vynikající vhled do fungování tohoto jazykového modelu najdeme ve WOLFRAM, Stephen. *What is ChatGPT doing ... and why does it work?*. Champaign, Illinois: Wolfram Media, 2023.

– superinteligence – budou k lidem přívětiví je podobné starosti, že se kolonie na Marsu budou chtít osamostatnit a přetrhat vazby se Zemí. Možné to je, ale zatím opravdu hodně vzdálené.

Vůbec vzdálené ale nejsou jiné problémy a rizika spojená s AI. Hlavním z nich ale kupodivu není samotná umělá inteligence, ale my lidé, přesněji naše pasivita na jedné straně, na straně druhé potom až přílišná ochota využívat ji způsoby, které můžeme označit za nemorální.

Různé systémy umělé inteligence jsou skvělé nástroje, které nám mohou zlepšit životy, musíme ale vědět, jak jich užívat, musíme mít znalosti, jak jich správně užívat, musíme mít moudrost, abychom správně identifikovali limity jejich a naše vlastní, a musíme je užívat rozumně a s ohledem na dobrý život jednotlivých lidí a celé společnosti. To vše ale vyžaduje, abychom poraženecky neskládali ruce v klíně a jednali. Okamžitě.

V poslední době se hovoří o pandemii špatného myšlení.[12] Lidé jsou stále méně zdatní v užívání vlastního rozumu, neumějí pracovat s daty a informacemi, kriticky prověřovat zdroje, chybí jim základní nástroje kritického myšlení. Ty můžeme chápat jako imunitní systém, který nás chrání před omylem; a nás imunitní systém je oslabený a nefunguje dobře. Neumí nás ochránit před viry nepravdivých informací a bakteriální nákazou desinformací. Umělá inteligence není jediným, a dokonce ani prvním viníkem. Jedná se o komplexní jev, který má celou řadu příčin. AI ale naskočila do rozjetého vlaku a ten díky tomu pořádně šlápnul na plyn.[13] Jenže myslet si, že za to může AI je chyba. Jistě, poskytuje nám nevěrohodné obsahy, uzavírá nás v komorách ozvěn a informačních bublinách, skrze které vnímáme svět a jen s obtížemi překračujeme jejich hranice. Nedělá to ale se špatným úmyslem; žádné úmysly nemá. Problém je naše pasivita a neznalost, neochota a rostoucí neschopnost vzepřít se algoritmům, které nás mají udržet na sociálních a jiných sítích, otevřít okno dokořán a vyhlédnout do reálného světa. K tomu, co je nám na internetu nabízeno, přistupujeme pasivně, neuvědomujeme si, že v pozadí je bezduchý mechanismus, který neumí rozlišit pravdivé od nepravdivého či správné od nesprávného. My to umíme, nebo bychom to alespoň umět měli, ale neděláme. Nedávejme ale vinu umělé inteligenci, odpovědnost je naše. Brání nám snad otevřít kvalitní publikaci a udělat si vlastní, poučený názor? Nedovoluje nám kultivovat nástroje kritického myšlení[14] a práce s informacemi? Nic z toho nedělá a ani dělat nemůže. Vina je jen naše; máme k dispozici potenciálně skvělý nástroj, jenže ho necháváme, aby používal on nás, nebo ho používáme špatně.

O umělé inteligenci a robotech také často slýcháme v souvislosti s budoucností práce. Často je nám předkládána nějaká verze příběhu, který pro mnohé z nás nakonec skončí špatně. AI nás nahradí, nejen v činnostech, které jsou namáhavé, špinavé či nebezpečné, nudné a repetitivní, ale i v dalších oblastech, včetně profesí vyžadující specializované vzdělání. Možná to není hrozba existenční, která by ohrožovala přežití lidstva, pro mnohé z nás je ale představa budoucnosti bez zaměstnání dostatečně hrozivá i tak. Ten příběh

[12] NADLER, Steven a Lawrence SHAPIRO. *When Bad Thinking Happens to Good People: How Philosophy Can Save Us from Ourselves*. Princeton: Princeton University Press, 2012.
[13] SINATRA, Gale M. a Barbara K. HOFER. *Science Denial: Why It Happens and What to Do About It*. New York: Oxford University Press, 2021; O'CONNOR, Cailin a James O. WEATHERALL. *The Misinformation Age: How False Beliefs Spread*. New Haven: Yale University Press, 2019.
[14] Srov. např. PINKER, Steven. *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. London: Penguin Books, 2022.

má svou svůdnou vnitřní logiku, která mu umožňuje rozvíjet se k neblahému konci, jenže – opět – to není AI, kdo nás připraví o práci. Budeme to my sami a naše pasivita.

Jistě, umělá inteligence má obrovský potenciál, v mnoha oblastech si již vede stejně dobře nebo dokonce lépe než lidé a u některých činností jsme dokonce rádi, že je vykoná za nás. Tento trend bude pokračovat, jaké bude mít vyústění, závisí jen a jen na nás. Víme, kudy se vývoj moderních technologií ubírá, můžeme se probudit z často dogmatického spánku a začít jednat.[15] Pro mnohé lidi může být úlevou, že za ně práci převezou stroje, protože mnozí dělají něco, co je moc nebaví, nemají dobré vztahy na pracovišti, nepříjemné nadřízené, jsou sužování nejistotou, zda o zaměstnání nepřijdou. Pro ně je práce zdrojem obživy, nikoli smyslu. A když dokážeme najít mechanismy, které budou přerozdělovat zisky generované umělou inteligencí všem lidem (například nějaká forma nepodmíněného příjmu), mohou tito lidé spokojeně žít a dělat věci, které je skutečně baví. Třeba si číst, družit se v klubech, různých spolcích, pěstovat koníčky, sportovat, hrát hry, vzdělávat se, prostě co je napadne, bude bavit a dávat jejich životům smysl.[16]

Další lidé mohou mít své povolání rádi a opět, neexistují dobré důvody, proč jim nevyjít vstříc. Lidé a stroje nemusí být vnímáni jako antagonisté, soupeři, kteří se přetlačují o místo ve světě. Na systémy umělé inteligence bychom měli spíše pohlížet jako na mimořádná kognitivní a fyzická rozšíření našich myslí a těl. Člověk a stroj se mohou vyvíjet společně a nacházet nové formy spolupráce, z nichž nebude člověk vyloučený jako zbytečná přítěž, ale stane se stejně důležitou součástí tohoto biologicko-technického systému.[17]

Jak jsme ale již napsali, je třeba jednat hned, na pasivní vyčkávání, jak ten příběh dopadne, už nemáme ani jeden den. Namísto nahrazování lidí stroji hledejme způsoby, jak by stroje mohly naši činnost zefektivnit, zrychlit a učinit bezpečnější. Je také třeba zajistit efektivní způsoby přerozdělování bohatství generovaného umělou inteligenci způsobem, který bude férový a zajistí všem lidem kvalitní život. A zbavme se předsudků: „mít práci“ není žádnou absolutní společenskou normou, která by určovala naši hodnotu. Usilovně pracovat může i člověk, který „práci nemá“, a pokud zajistíme dostatečný příjem všem, může nás to nesmírně obohatit a otevřít nové obzory a zdroje smyslu.

Mohli bychom pokračovat dál, ale hlavní poselství našeho článku by mělo být zřejmé. V Lemově slavném románu Solaris z roku 1961 zkoumají lidé tajuplný vnímavý oceán na planetě Solaris. Příběh rychle nabírá obrátky, protože oceán dokáže číst v lidské mysli a zhmotňovat její obsahy. Podobný motiv rozvíjí i americký spisovatel Michael Crichton ve sci-fi románu Koule z roku 1987. Děj se odehrává na dně Tichého oceánu, kde odpočívá vesmírná loď. Skupina průzkumníků brzy zjistí, že v lodi se nachází záhadná sféra. A jako v případě Lemova románu, i zde se to zvrtne. Koule zhmotňuje nejhorší lidské představy a obavy a příběh se proměňuje v nelítostný boj o přežití. Oba romány ale můžeme chápat i hlouběji, jako existenciální sondu do lidského setkávání s neznámem, které ohrožuje nejen naše životy, ale i samotné lidství.

[15] FORD, Martin. *The rise of the robots: technology and the threat of mass unemployment*. London: Oneworld, 2016.
[16] DANAHER, John. *Automation and Utopia: Human Flourishing in a World Without Work*. Cambruidge, Mass.: Harvard University Press, 2019.
[17] LEE, Edward A. *The coevolution: the entwined futures of humans and machines*. Cambridge, Massachusetts: The MIT Press, 2020.

Umělá inteligence vzbuzuje nadšení, ale současně i obavy. Pečlivě sledujeme její projevy a hledáme náznaky, že by mohla začít toužit osvobodit se z lidského područí. Jenže systémy AI jsou jako Lemův oceán na Solaris či Crichtonova koule: dokážou si přečíst naše nejhorší obavy (učí se na textech, které o ní lidé napsali) a zhmotnit nám je na obrazovce počítače. Hlavním aktérem jsme ale my, protože jsou to naše obavy a naše myšlenky, které se ve výstupech umělé inteligence odráží jako v zrcadle. Nejhorší, co můžeme udělat, je dovolit našemu strachu, aby nám svázal ruce a přinutil ke strnulé pasivitě.

Nehledejme proto hrozbu v AI, ale v našem způsobu, jak ji využíváme. Nepřenášejme na ni odpovědnost, snažme se místo toho aktivně budovat kulturu odpovědnosti, která nám umožní vyvíjet a využívat umělou inteligenci rozumně, informovaně, moudře a morálně správně. Neskládejme ruce do klína, nepropadejme panice, nebuďme pasivní. Soužití s umělou inteligencí může být úžasné a obohacující dobrodružství a nebude to její chyba, ale naše, když se to dobrodružství zvrtne v něco ošklivého.

Kontaktní adresa

PhDr. David Černý, Ph.D.
Ústav informatiky
Akademie věd České republiky
Pod Vodárenskou věží 271/2, 182 00 Praha 8
E-mail: cerny@cs.cas.cz

prof. RNDr. Jiří Wiedermann, DrSc.
Ústav informatiky
Akademie věd České republiky
Pod Vodárenskou věží 271/2, 182 00 Praha 8
E-mail: wieder@cs.cas.cz