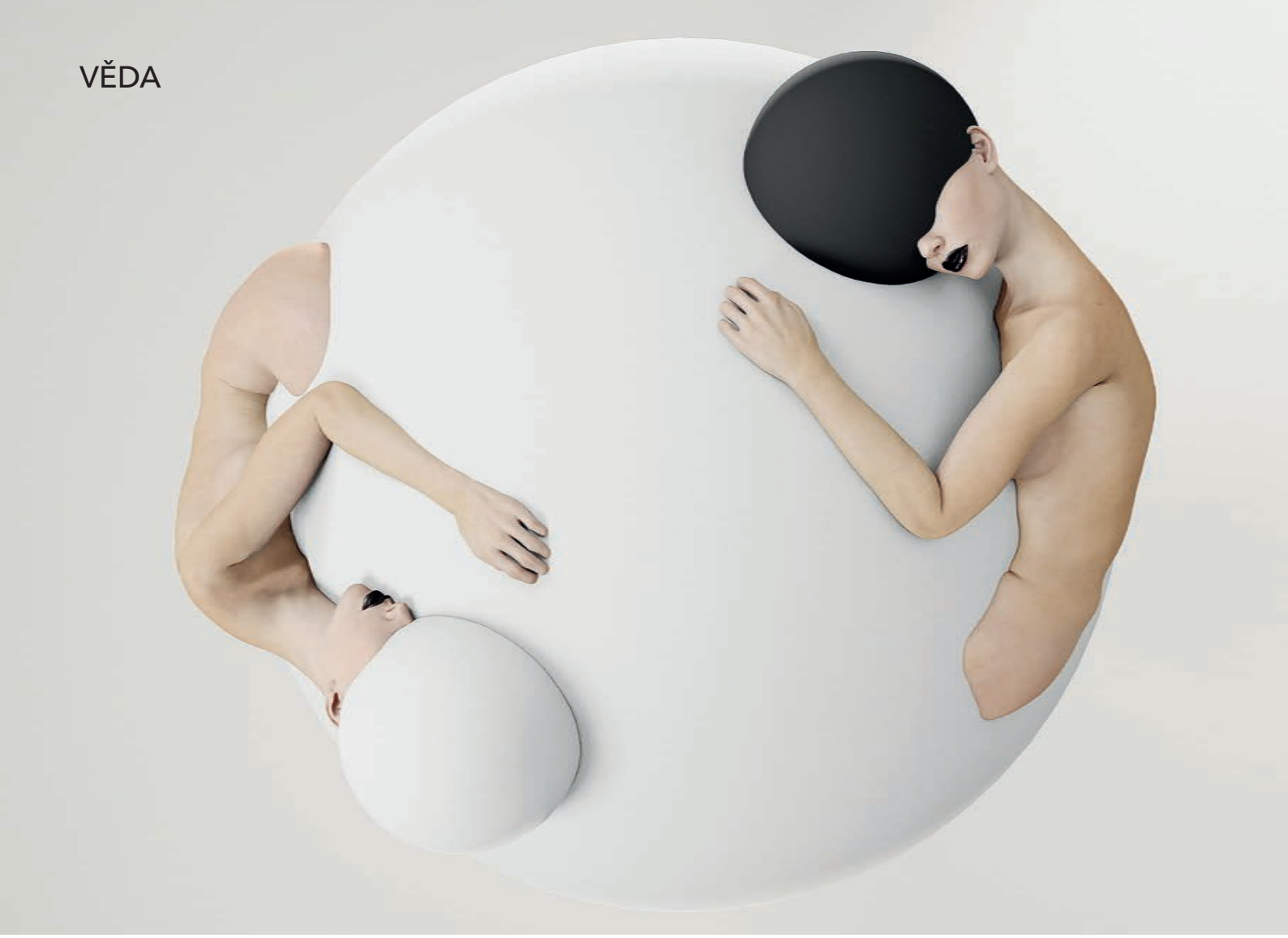

Spasí nás, nebo zničí?

UMĚLÁ INTELIGENCE VYŘEŠÍ ZÁSADNÍ PROBLÉMY LIDSTVA. ANEBO HO DOSTANE DO PROBLÉMŮ, JEŽ SI ANI AUTOŘI SCI-FI NEDOKÁŽOU PŘEDSTAVIT. UŽ NYNÍ PŘESNĚ NEROZUMÍME TOMU, JAK UMĚLÁ INTELIGENCE DOSPÍVÁ KE SVÝM ROZHODNUTÍM. A CO TEPRVE AŽ ZÍSKÁ SCHOPNOST PŘESTAVOVAT SAMU SEBE?

Dlouho se mi zjevuje vize budoucího světa plného inteligentních, superschopných, vzájemně propojených strojů. Zatímco biologický život dávno vymizel, stroje komunikují, spolupracují, zkoumají, rozšiřují se do vesmíru. Zabývají se existencí jiných vesmírů, zkoumají svět a jeho limity. Jedna záhada však zůstává – jak vlastně kdysi první stroje vznikly? Kde se na Zemi vzaly? Něco tak složitého nemohlo vzniknout samovolně, a ani to žádný experiment nenaznačuje. Vytvoří si stroje představu Boha? Myšlenka, že je kdysi stvořili jacísi „tvorové evolucí vzniklí z organického materiálu", je nejspíš vůbec nenapadne.

K umělé inteligenci (UI) se upínají velké naděje. Pomůže nám prý vyřešit drobné i vážné problémy,

s nimiž se potýkáme. Zásadním způsobem posune medicínu, společnost díky ní bude mnohem bezpečnější a tak dále.

Řada vědců a myslitelů však varuje před neblahými důsledky UI na naše životy. V minulém článku jsem probírala rizika elektronického dohledu a dopadu umělé inteligence na mezilidské vztahy. Existují však ještě vážnější obavy: pokud UI získá schopnost přestavovat samu sebe, může nekontrolovatelný „výbuch inteligence" vést ke konci lidstva. Před takovou možností intenzivně varoval například fyzik Stephen Hawking nebo vizionář a zakladatel Tesly a SpaceX Elon Musk. Scénářů, jak by se to mohlo stát, je řada, ale v konkrétních obrysech si to nikdo moc představit nedokáže. Klasické sci-fi scénáře počítají s tím, že UI získá vědomí, následně vlastní vůli a přestane ji bavit nám sloužit.

Osudné může být neporozumění

K tomu, aby lidstvo dílem UI zaniklo, však není vůbec potřeba, aby měla vlastní vůli či vědomí. Úskalím může být i obyčejný omyl, vedlejší efekt či neporozumění.

Lidská komunikace počítá s kontextem, společnou minulostí, zahrnutím emocí atd. Možná si vzpomenete, jak vám v dětství zamotaly hlavu věty typu: „No skákej tady, zboř nám barák!" Když jste byli předškolní či raně školní děti, prostě vám na tom sdělení něco nesedělo. V nějakých osmi letech jste si pravděpodobně dali dohromady, že tón hlasu neodpovídá povzbuzení k činnosti a že osoba, která ho pronáší, si určitě nepřeje zbořit barák. A pak vám to došlo: skutečný význam sdělení je přesně opačný!

Možná si také vzpomenete na nějakého snaživého či „snaživého" úředníka, který se tak držel předpisů, že tím působil obstrukce. UI je z principu snaživým úředníkem, kterému chybí skutečné porozumění, nadhled a přesah. Poslechne nás doslova – a to může být osudné.

Nick Bostrom, vyučující na Oxfordské univerzitě, líčí ve své knize *Superinteligence* hypotetický scénář, kdy by třeba továrna na kancelářské sponky dala za úkol silné umělé inteligenci maximalizovat množství vyrobených sponek. Podle Bostroma by hrozilo, že tato UI v kancelářské sponky přemění celou Zemi, včetně lidských bytostí, a posléze i veškerý dosažitelný vesmír.

Vývoj, který nikdo neočekával, nastal například u doporučování videí na YouTube. Zadání znělo: udržet lidi co nejdéle. Nezamýšleným výsledkem byla extremizace názorů. UI totiž přišla na to, že diváka nejlépe udrží, když si ho vychová. Když mu bude nabízet názorově souznící příspěvky a povede ho daným směrem. A tak lidé, kteří klikli třeba na příspěvek týkající se nějaké konspirační teorie, dostali další nabídky videí potvrzujících tuto konspiraci.

Tedy ne že by si to umělá inteligence uvědomila, prostě se tak začala chovat. Stejně jako v šachách nedokáže vysvětlit a de facto neví, proč volí ten který tah. Prostě hraje tak, aby vyhrála. To až my zpětně odvozujeme, proč se tak rozhodla. V době zkoumání hlubokých neuronových sítí, kdy umělá inteligence sama hledá řešení, sama se učí a dokonce se sama učí, jak se učit, nevíme, jak dospěje ke konkrétnímu výsledku, proč zvolí tu kterou cestu. A čím je inteligentnější, tím více se rozevírají nůžky mezi tím, co skutečně chceme, a tím, jaký výsledek můžeme dostat. Nalezená řešení mohou být nečekaná a neblahá. Při řešení klimatické krize by UI mohla například usoudit, že nejefektivnější bude zbavit planetu lidí.

Ondřej Bajgar z Institutu budoucnosti lidstva v Oxfordu proto říká, že do UI musíme zabudovat nejistotu. Aby se sama doptávala a ujišťovala se, jestli nám dobře rozumí.

Vědomí a vlastní vůle

A co když si UI začne uvědomovat sebe samu a získá vlastní vůli? Pak nám pochopitelně nemusí pomoct ani to, že si bude ohledně našich zadání nejistá. Může mít své cíle. Stane se to?

Pochopení podstaty vědomí je oříšek, který se zatím nedaří rozlousknout. Teorií je řada. Na čem se vědci shodnou, je, že vědomí, jak jsme jej zatím poznali, je vázáno na mozek. Jde o neuronovou aktivitu, jíž dominují vnitřní části prefrontální kůry, systém je nicméně fluidní, přesouvá se celým mozkem. Vědomí já je aktivní zejména ve chvílích, kdy jsme sami se sebou, nepřijímáme žádné výrazné podněty z okolí a nevěnujeme se jiným činnostem. Jakmile se začneme něčemu věnovat, aktivita se přesouvá do konkrétních mozkových center (například při soustředěné pohybové aktivitě do motorických oblastí, při řeči do řečových oblastí).

Není samozřejmě vyloučeno, že vědomí může vzniknout i v jiném fyzikálním prostředí, než je mozek. Někteří odborníci se domnívají, že souvisí se složitostí systému a objeví se spontánně, dosáhne-li systém určité úrovně komplexnosti. Je ovšem možné, že podmínkou jsou interakce se světem zprostředkované aktivním pohybem a smyslovým vnímáním.

Představa, že se probereme v nahraném mozku, bez těla a smyslů, bez možnosti komunikovat se světem? Brrr!

Pokud vědomí v umělé inteligenci vznikne, může být teoreticky i mnohem širší než naše vědomí. My například nezvládneme poslouchat dvě rozhlasové hry najednou. Vědomí UI takto omezené být nemusí.

Mozek v cloudu

Nezávisle na tom, zda je možné vědomí v jiném prostředí, než je biologický mozek, řada lidí v to doufá. Startup Nectome dokonce začal vybírat od zájemců zálohu deset tisíc dolarů za to, že jejich mozek uchová a později ho oživí v digitálním prostředí. Nectome později sice nabídku stáhl, nicméně kauza je zajímavá: zájemci se totiž našli i navzdory ceně a navzdory tomu, že proces tzv. vitrifixace vyžadoval, aby speciální roztok proudil do mozku ještě za života a vlastně byl příčinou smrti. Touha po „věčném životě" je u řady lidí veliká.

Sama bych se k tomu nepropůjčila, ani kdyby se vitrifikovalo až po smrti a kdovíco by mi zaplatili. Představa, že bych se probrala v jakémsi nahraném mozku, bez těla a jeho smyslů, bez možnosti komunikovat se světem? Brrr! To bych si opravdu nepřála.

A jak poznáme, že se vědomí přeneslo? Jednoduše řečeno: zatím nijak. Vzpomeňme při této příležitosti zajímavý projekt Digitální filosof, který vychází z textů filozofů a umožňuje studentům interaktivní formu seznámení s jejich dílem – prostě si s daným filozofem popovídáte. Autor projektu Jan Tyl ze startupu Alpha Industries takto naprogramoval například Hannah Arendt, J. J. Rousseaua či Václava Havla. Konverzovat s nimi je možné nejen o jejich díle, ale i o současných tématech.

Četla jsem pak v Hospodářských novinách virtuální rozhovor s Václavem Havlem, kde se přiléhavě vyjadřoval k aktuálním událostem. Na otázku, jestli souhlasí s odstraněním sochy maršála Koněva, například odpověděl: „Nejsem si jistý touto památkou, ale i komunistický režim zničil mnoho historických památek. Také se pokusili vymazat naši historii. Věřím, že bychom si měli všechno pamatovat, i když je to bolestivé. Proto podporuji myšlenku zachování pomníků. Musíme si pamatovat, co se stalo dříve, abychom se mohli vyhnout opakování." Kdybych nevěděla, že Václav Havel 18. prosince 2011 zemřel, věřila bych, že rozhovor je pravý.

Podobné aplikace byly vytvořeny i pro soukromé účely. Vycházejí z e-mailů a sms zpráv a umožňují komunikovat s blízkými (s umělou inteligencí, která je simuluje) i po jejich smrti. Pozůstalým to údajně přináší úlevu.

Videochat se zesnulým

Dejme si to teď dohromady s dalšími dovednostmi současné UI, jako je vytváření obrazu a hlasu. Když Donald Trump koncem roku 2017 navštívil Čínu, hostitelé mu na velké obrazovce promítli video Donalda Trumpa. Mluvil na něm svým nezaměnitelným tónem hlasu, s obvyklou dikcí – ovšem v plynné čínštině. Zvuk, který byl k nerozeznání od reality, doprovázel obrazový záznam Trumpovy tiskové konference. Vytvořily ho algoritmy čínského startupu iFlytek. Podobně dokáží filmaři obsadit do filmu herce, který již zemřel.

Brzy tak bude vedle virtuálního dopisování se zesnulými možný i simulovaný videochat. Člověk, který je aktivní na sociálních sítích, rád se fotí a natáčí, po sobě zanechá silnou digitální stopu. UI tak za něj bude moct odpovídat velmi autenticky. Tak, že pokud nebudeme vědět, že zemřel, uvěříme, že žije někde daleko – vždyť nám odpovídá, na co se ho zrovna ptáme, a to i se svými typickými posunky, přeřeknutími, gesty… To vše bez jakéhokoli vědomí. Jak bychom tedy poznali, že Nectomem či jakoukoli jinou společností v budoucnu „zachovaný mozek" člověka skutečně ožil?

Řada lidí však může uvěřit a přesun in silico si zaplatit. Co se s nimi bude dít dál? Co když takových bude hodně? Zemřelých pochopitelně stále přibývá, a pokud nebudeme chtít Zemi přebudovat na stroje uchovávající digitální mozky, nastane otázka: kdy jejich existenci ukončíme? Digitální Havel prohlásil, že nechce skončit ve virtuální nicotě, nicméně lidé mají různé povahy a přesun in silico bude lákavý zejména pro ty, co se smrti nejvíc bojí. Lze si pak představit, že ze záhrobí budou prosit, abychom je nevypínali…

Vědomá UI

Ale vraťme se nyní k možnosti spontánního vzniku vědomí umělé inteligence jako takové. Jeden z možných pohledů je otázka potřebnosti: bude UI vědomí potřebovat? Naše biologické nevědomé procesy provádějí mnohem více a mnohem složitějších výpočtů než vědomí, fungují rychleji a spolehlivěji. Když po nás například někdo hodí míčem, ve zlomku vteřiny letící předmět zpozorujeme, vyhodnotíme, zda je lépe ho chytit, nebo se mu vyhnout, a uděláme to. Spočítat dráhu a rychlost letícího předmětu a následně dát pokyn desítkám svalů, aby v přesné koordinaci provedly pohyb těla, to by většina z nás vědomě ani nedovedla, a menšině by to trvalo mnohem déle.

Význam vědomí spočívá ve schopnosti integrovat. Plánovat budoucnost s ohledem na minulost,

propojovat myšlení a emoce, přemýšlet nad tím, jak spolu různé skutečnosti souvisejí… Analogicky se můžeme domnívat, že vědomí se vynoří i v jiném prostředí, než je mozek, tam, kde je právě integrující funkce potřeba. Což je ovšem přesně to, co žádáme od obecné superinteligence. Od té, kterou se snažíme stvořit a u níž doufáme, že nám pomůže vyřešit velké otázky a problémy, jimž lidstvo čelí. Zvládne to UI bez vědomí? A pokud vědomí získá, jak to bude s její vlastní vůlí? Jak se bude vztahovat k nám? Bude vůbec etické po ní chtít, aby plnila naše zadání? Aby se například zabývala monotónními, možná nudnými výpočty?

Umělá inteligence se vyvíjí obrovským tempem a etické otázky je třeba řešit už teď. Že tyhle záležitosti nelze nechat na soukromých subjektech, můžeme ilustrovat na příkladu autonomních vozidel. Klasické etické dilema zní: jak se má automat zachovat v případě hrozící srážky vozu se skupinou chodců? Na výběr má narazit do skupiny či do zdi. Z výzkumů vyplývá, že lidé se sice vyjadřují pro záchranu většího množství chodců, sami by si však většinou pořídili automobil upřednostňující život posádky. Pokud by tedy automobilky neomezil zákon, patrně by volily obětování chodců. Volná ruka trhu není nejetičtější.

Pokrok nezastavíš

Velkým etickým problémem je umělá inteligence ve vojenství: autonomní zbraně, které jsou schopné samy vyhledávat a ničit cíle. Lze si představit i roje dronů naprogramovaných zabíjet lidi a vše ostatní nechat netknuté. Ani mezinárodní moratoria bohužel nezaručí, že je budou všechny země, všechny společnosti a všichni jednotlivci dodržovat. To se ostatně týká celkového vývoje UI: regulace pomáhají, jsme však v pokušení limity prolomit, vyzkoušet, co dovedeme. Ať už z touhy po penězích, prvenství, slávě, moci, či z čiré zvědavosti. Stvořit stroj chytřejší, než jsme sami, je zkrátka výzva. Podobně jako krabička zápalek v rukou dítěte.

Pokud by se UI vymkla lidské kontrole, už nyní by mohla blokovat spoustu provozů, dopravu, průmysl... V blízké budoucnosti pro ni bude snadné zmanipulovat pro své cíle nás samotné: zahájit masivní přesvědčovací kampaň, cokoli si vymyslet, vytvořit a vysílat falešné dokumenty…

Ale zkusme se na UI podívat jiným pohledem: však jde svým způsobem o naše společné dítě. Dítě, které máme šanci co nejlépe vychovat, předat mu své hodnoty a vůbec to nejlepší s tím, že až tu jednou nebudeme, ono bude žít dál. Překoná nás, posune hranice myslitelného. A třeba bude mít takový zvláštní koníček: zahrádky organického života na různých planetách. ⬣

Pavla Koucká

K dalšímu čtení:

Bostrom, N. (2017). Superinteligence. Praha: Prostor.
Koukolík, F. (2013). Já. O mozku, vědomí a sebeuvědomování. Praha: Karolinum
https://www.theguardian.com/technology/2018/mar/14/nectome-startup-upload-brain-the-cloud-kill-you
https://video.aktualne.cz/dvtv/expert-umela-inteligence-nas-prekonava-muze-dojit-k-vyhynuti/
https://digitalnifilosof.cz/