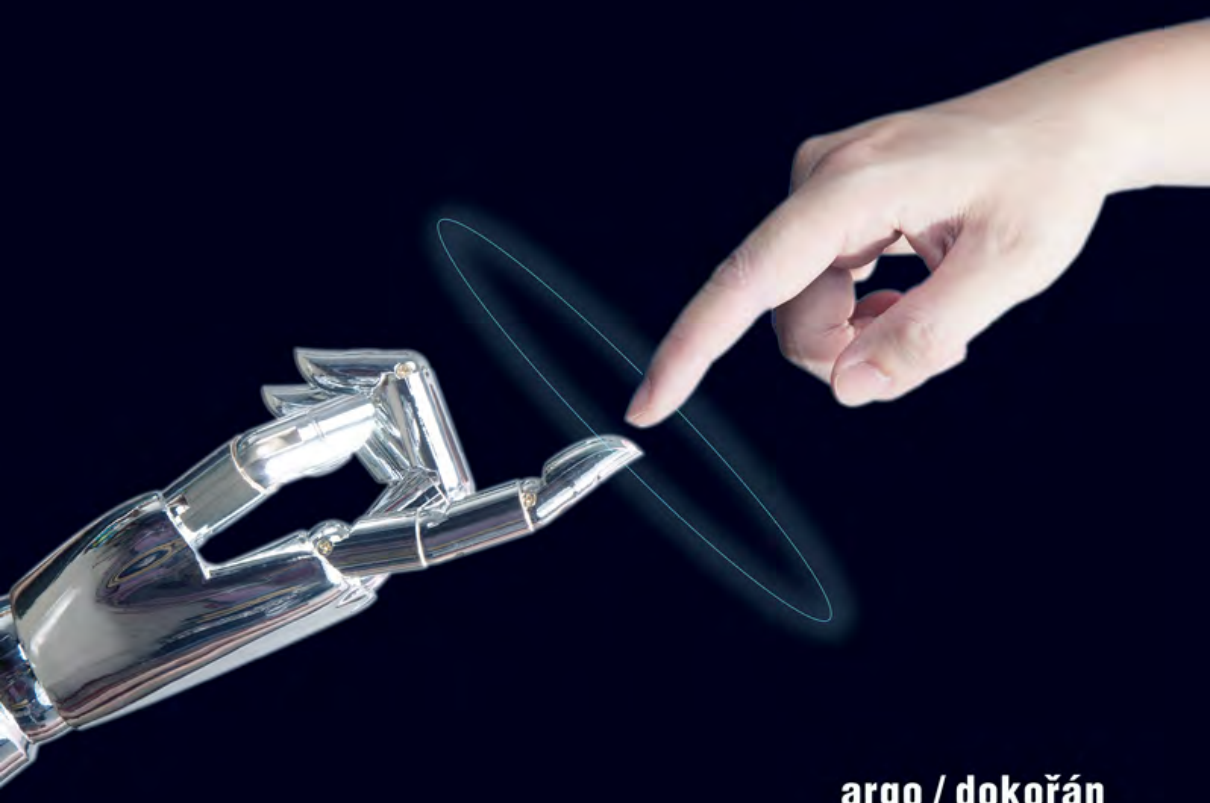
Stuart Russell

JAKO ČLOVĚK

Umělá inteligence a problém jejího ovládání



argo / dokořán

Stuart Russell

JAKO ČLOVĚK

Umělá inteligence a problém jejího ovládání

ARGO / DOKOŘÁN

Stuart Russell

Jako člověk

Umělá inteligence a problém jejího ovládání

Druhé vydání v českém jazyce (první elektronické).
Z anglického originálu *Human Compatible: Artificial Intelligence and the Problem of Control* přeložil Jiří Zlatuška.
Odpovědný redaktor Zdeněk Kárník.
Redakce Klára Soukupová.
Obálka podle návrhu Pavla Růta, sazba a konverze do elektronické verze Michal Puhač.
Vydalo v roce 2021 nakladatelství Dokořán, s. r. o.,
Holečkova 9, Praha 5,
dokoran@dokoran.cz, www.dokoran.cz,
jako svou 1134. publikaci (367. elektronická).

ISBN 978-80-7675-043-2

OBSAH

PŘEDMLUVA

PROČ TATO KNIHA? PROČ NYNÍ?

Tato kniha pojednává o minulosti, současnosti a budoucnosti našich snah pochopit a vytvořit inteligenci. Jde o důležité téma nikoli proto, že se umělá inteligence (AI) rychle stává všeprostupujícím aspektem současnosti, ale proto, že jde o dominantní technologii budoucnosti. Světové mocnosti si této skutečnosti začínají všímat a největší světové korporace to již nějakou dobu vědí. Nedokážeme přesně předpovědět, jak se tato technologie bude vyvíjet ani jak rychle. Musíme se nicméně připravovat na možnost, že se stroje v rozhodování stanou daleko schopnějšími než lidé. A co pak?

Vše, co může civilizace nabídnout, je produktem naší inteligence; kdybychom získali přístup ke značně vyšší inteligenci, znamenalo by to největší událost v historii lidstva. Cílem této knihy je vysvětlit, proč by se mohlo jednat o událost v historii lidstva poslední a co dělat, abychom takovým koncům dokázali zabránit.

PŘEHLED

Kniha sestává ze tří částí. První část (kapitoly 1–3) zkoumá ideu inteligence u lidí a u strojů. Tento výklad nevyžaduje žádné technické znalosti, ale pro zájemce je doplněn čtyřmi přílohami, jež vysvětlují některé z ústředních pojmů v základech současných systémů AI. Druhá část (kapitoly 4–6) je věnována diskusi o některých problémech vycházejících z toho, že stroje jsou nadány inteligencí. Zvláště se soustředíme na problém ovládání: zachování absolutní moci nad stroji, které budou mocnější než my. Třetí část (kapitoly 7–10) naznačuje nový způsob uvažování o AI a o zajištění, že stroje zůstanou navždy pro člověka prospěšné. Kniha je určena širokému publiku, ale doufám, že i přivede odborníky na umělou inteligenci k tomu, aby znovu promýšleli své základní předpoklady.

KAPITOLA 1.

POKUD USPĚJEME

Mí rodiče žili před mnoha lety v anglickém Birminghamu v domě poblíž univerzity. Rozhodli se odstěhovat mimo město a dům prodali Davidu Lodgeovi, profesoru anglické literatury. Ten už byl tou dobou dobře známým spisovatelem. Nikdy jsem se s ním nesetkal, ale přečetl jsem některé z jeho knih: *Hostující profesoři* a *Svět je malý*. Hlavními postavami jsou v nich smyšlené postavy akademiků, kteří se stěhují z fikčního Birminghamu do fikčního kalifornského Berkeley. Jako skutečný akademik ze skutečného Birminghamu, jenž se právě přestěhoval do skutečného Berkeley, jsem měl pocit, že se mne někdo z *katedry shod okolností* snaží na něco upozornit.

Zvlášť mne zasáhla jedna konkrétní scéna ze *Svět je malý*: hrdina knihy, snaživý literární teoretik, se zúčastní významné mezinárodní konference a pokládá dotaz panelu vůdčích vědců: „Co se stane, pokud s vámi bude každý souhlasit?“ Tato otázka vyvolá zděšení, protože se tito panelisté zajímají víc o intelektuální souboje než o prokazování pravdy nebo dosahování pochopení. Přišlo mi tehdy, že by se podobná otázka měla pokládat vůdčím osobnostem v oboru umělé inteligence: „Co když uspějete?“ Hlavním úkolem této disciplíny vždycky bylo vytvořit AI lidské nebo nadlidské úrovně, ale bez jakékoli starosti o to, co se stane, když se to podaří.

Několik let poté jsem s Peterem Norvigem začal pracovat na nové učebnici AI, jejíž první vydání vyšlo v roce 1995.[1] Poslední kapitola této knihy se jmenuje „Co když opravdu uspějeme?“. Ukazuje možnosti dobrých i špatných výsledků, ale nedochází k žádnému pevnému závěru. V době třetího vydání roku 2010 už konečně mnozí začali uvažovat nad možností, že nadlidská AI nemusí být přínosem – ale jednalo se převážně o lidi mimo AI, nikoli o vědce hlavního proudu v ní. V roce 2013 jsem dospěl k přesvědčení, že tento problém nejen do hlavního proudu patří, ale že se jedná možná o nejdůležitější otázku, před kterou lidstvo stojí.

V listopadu 2013 jsem měl přednášku v Dulwich Picture Gallery, respektovaném muzeu umění v jižním Londýně. V publiku byli převážně penzisté – lidé, kteří nebyli vědci, ale měli obecný zájem o intelektuální problémy –, a tak jsem musel svou přednášku pojmout bez jakýchkoli technických aspektů. Připadlo mi to jako místo vhodné k tomu, abych si poprvé vyzkoušel své myšlenky před veřejností. Poté, co jsem vysvětlil, čeho se AI týká, uvedl jsem své návrhy na pět kandidátů na „největší událost v budoucnosti lidstva“:

1. Všichni zemřeme (dopad asteroidu, klimatická katastrofa, pandemie a podobně).
2. Budeme žít věčně (medicína zvládne stárnutí).
3. Vynalezneme cestování rychlostmi většími než světlo a podmaníme si vesmír.
4. Navštíví nás nadřazená cizí civilizace.
5. Vynalezneme superinteligentní AI.

Uvedl jsem, že pátý kandidát, superinteligentní AI, by mohl být vítězem, protože by nám pomohl odvrátit fyzické katastrofy a umožnil dosáhnout věčného života i cestování rychlostmi většími než světlo, pokud by to vše bylo skutečně možné. Znamenalo by to obrovský skok – diskontinuitu – v naší civilizaci. Příchod superinteligentní AI je z mnoha pohledů analogický příchodu mimozemské civilizace, ale mnohem pravděpodobnější. A co je snad nejdůležitější, AI je na rozdíl od mimozemšťanů něco, co můžeme ovlivnit.

Pak jsem toto publikum požádal, aby si představili, co by se stalo, pokud bychom dostali zprávu od nějaké mimozemské civilizace, že během třiceti až padesáti let dorazí na Zemi. Slovo *vřava* se k popisu jejich reakce moc nehodí. Naše odpověď na předpokládaný příchod superinteligentní AI je... hm, *nedostatek zájmu* je výraz, který ji začíná popisovat. (V pozdější přednášce jsem situaci ilustroval v podobě e-mailové komunikace z obrázku 1.) Vysvětlil jsem nakonec význam superinteligentní AI takto: „Úspěch by byl největší událostí v lidské historii... a možná také událostí v lidské historii poslední."

O několik měsíců později, v dubnu 2014, jsem byl na konferenci na Islandu, kam mi volali z pořadu v National Public Radio a ptali se, jestli by se mnou mohli

Od: Nadřazená mimozemská civilizace <sac12@sirius.canismajor.u>
Komu: lidstvo@UN.org
Předmět: Kontakt
Varujeme vás: přijdeme během 30–50 let.

Od: lidstvo@UN.org
Komu: Nadřazená mimozemská civilizace <sac12@sirius.canismajor.u>
Předmět: Out of office: Re: Kontakt
Lidstvo zrovna není v práci. Ozveme se vám na vaši zprávu, až se vrátíme. :)

Obr. 1: Málo pravděpodobná výměna e-mailů následující po prvním kontaktu nadřazené mimozemské civilizace.

udělat rozhovor o filmu *Transcendence*, který měl zrovna premiéru ve Spojených státech. Shrnutí děje a recenze na tento film jsem sice četl, ale neviděl jsem ho, protože jsem tou dobou žil v Paříži, kde ho uvedli až v červnu. Přidal jsem si ovšem tehdy na cestě z Islandu zpět zastávku v Bostonu, abych se mohl zúčastnit schůzky na ministerstvu obrany. Takže když jsem přiletěl na bostonské Loganovo letiště, vzal jsem to taxíkem do nejbližšího kina, kde tento film promítali. Sedl jsem si do druhé řady a sledoval, jak profesora AI z Berkeley, hraného Johnnym Deppem, odstřelí anti-AI aktivista, který se obává – jak jinak – superinteligentní AI. Bezděčně jsem se v sedadle schoulil. (Další upozornění z *katedry shod okolností*?) Před smrtí je vědomí postavy představované Johnnym Deppem nahráno do kvantového superpočítače, načež rychle překonává lidské schopnosti a hrozí, že se zmocní světa.

Dne 19. dubna 2014 vyšla v *The Huffington Post* recenze *Transcendence*, kterou napsali společně Max Tegmark, Frank Wilczek a Stephen Hawking. Byla v ní věta z mé přednášky v Dulwich o největší události v lidské historii. Od té chvíle jsem začal veřejně zastávat postoj, že má oblast výzkumu znamená potenciální riziko pro můj vlastní živočišný druh.

JAK JSME SEM DOŠLI?

Kořeny AI sahají daleko do minulosti, ale její „oficiální" počátek nastal v roce 1956. Dva mladí matematici, John McCarthy a Marvin Minsky, přesvědčili Clauda Shannona, tehdy již proslulého svými objevy v teorii informace, a Nathaniela Rochestera, tvůrce prvního komerčního počítače firmy IBM, aby se k nim přidali a společně zorganizovali letní školu na Dartmouth College. Jejich cíl zněl takto:

> Vycházíme z myšlenky, že každý aspekt učení nebo jakýkoli jiný rys inteligence lze principiálně popsat natolik přesně, že bude možné sestavit stroj, který ji bude simulovat. Pokusíme se zjistit, jak naučit stroje používat jazyk, vytvářet abstrakce a pojmy, řešit druhy problémů, které jsou nyní výhradně lidskou doménou, a zlepšovat se. Domníváme se, že v jednom či několika z těchto problémů může být učiněn značný pokrok, stačí, když se pečlivě vybraná skupina vědců sejde a budou spolu přes léto pracovat.

Je asi zbytečné konstatovat, že to trvalo mnohem déle než jedno léto: na těchto problémech se pracuje dodnes.

Někdy zhruba deset let po dartmouthském setkání měla AI několik velkých úspěchů, mezi něž patří algoritmus Alana Robinsona pro obecné logické vyvozování[2] a program hrající dámu od Arthura Samuela, který se naučil porážet

svého tvůrce.[3] První bublina AI praskla na konci 60. let, když se nepodařilo splnit naděje vkládané do prvních snah o strojové učení a strojový překlad. Zpráva vypracovaná pro britskou vládu v roce 1973 došla k závěru: „V žádné, ani dílčí, oblasti zde objevy dosud dosažené nedošly ke slibovaným výrazným dopadům."[4] Jinými slovy, stroje prostě nebyly dost chytré.

Mé jedenáct let staré já bohužel o této zprávě nevědělo. Dva roky poté, když jsem dostal programovatelnou kalkulačku Sinclair Cambridge, jsem ji chtěl jen učinit inteligentní. S programem o délce nejvýše čtyřicet šest stisknutí tlačítka ovšem Sinclair nebyl pro inteligenci lidské úrovně dostatečný. Neodradilo mne to, a když jsem získal přístup ke gigantickému superpočítači CDC 6600[5] na Imperial College, napsal jsem program na hraní šachů – štos děrných štítků 60 cm vysoký. Nebyl moc dobrý, ale to nevadilo. Věděl jsem, co chci dělat.

V polovině 80. let jsem se stal profesorem na Berkeley a AI prožívala velký comeback díky komerčnímu potenciálu takzvaných expertních systémů. Druhá AI bublina praskla, když se ukázalo, že tyto systémy nestačily na mnoho z úkolů, na něž byly použity. Stroje opět nebyly dost chytré. Přišel zimní spánek AI. Na můj kurz na Berkeley, který nyní navštěvuje více než devět set studentů, docházelo v roce 1990 pouze dvacet pět zájemců.

Komunita kolem AI se z toho poučila: chytrost přístrojů se očividně zlepšila, ale museli jsme nejdříve udělat své domácí úlohy, aby k tomu došlo. Celá disciplína se stala více matematickou. Navázala se spojení s etablovanými disciplínami, jako jsou pravděpodobnost, statistika a teorie řízení. Semínka dnešního úspěchu AI byla zaseta během této zimy, včetně počátků práce na rozsáhlých pravděpodobnostních systémech vyvozování a toho, co později začalo být známo jako *hluboké učení*.

Kolem roku 2011 zažily techniky hlubokého učení dramatický pokrok v rozpoznávání řeči, v rozpoznávání vizuálních objektů a ve strojovém překladu – tedy ve třech z nejdůležitějších otevřených problémů celé oblasti. V závislosti na zvoleném srovnání se nyní stroje v těchto oblastech vyrovnají lidským schopnostem, nebo je předčí. V letech 2016 a 2017 systém AlphaGo z produkce DeepMind porazil Leeho Sedola, bývalého světového mistra ve hře go, a Ke Jieho, současného mistra – což jsou události, o nichž někteří experti předpovídali, že nemohou nastat dříve než v roce 2097, pokud vůbec.[6]

V současné době se AI objevuje na titulních stranách médií téměř každý den. Vznikly tisíce start-upů, které jsou poháněny přívalem rizikového kapitálu. Miliony studentů se zapsaly do online kurzů AI a strojového učení a experti v této oblasti berou platy v milionech dolarů ročně. Investice valící se z rizikových fondů, státních rozpočtů a velkých společností představují ročně desítky miliard dolarů – což za posledních pět let dělá víc než za celou předchozí historii disciplíny. Pokrok, který přichází – jako samořiditelná auta a inteligentní

osobní asistenti –, bude mít velmi pravděpodobně během další zhruba dekády podstatný dopad na svět. Potenciální ekonomické a společenské přínosy AI jsou ohromné, což vytváří nesmírný spád pohánějící výzkumné podnikání v AI.

CO SE STANE DÁL?

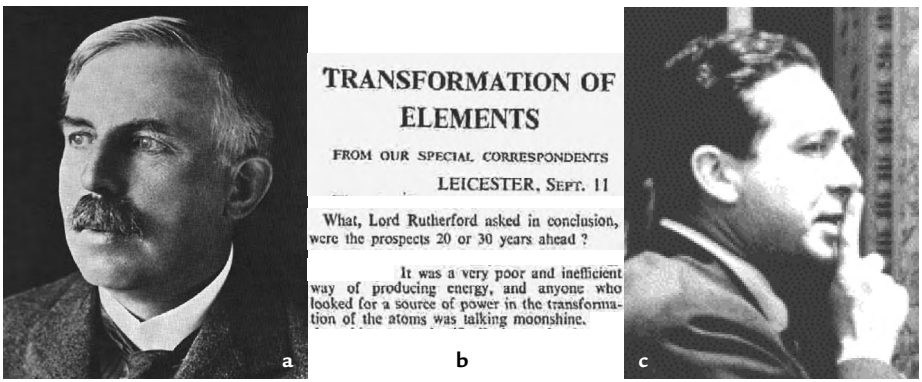
Znamená takto rychlé tempo pokroku, že budeme za chvíli předstiženi stroji? Ne. Je zde několik průlomových událostí, které se musí stát, než budeme mít cokoli, co by připomínalo nadlidskou inteligenci.

Průlomy ve vědě je vždy těžké předpovídat. Abychom si ukázali, jak těžké to je, podívejme se na historii jiné vědní oblasti s potenciálem ukončit naši civilizaci: jaderné fyziky.

V prvních letech dvacátého století zřejmě nebylo uznávanějšího jaderného fyzika, než byl Ernest Rutherford, objevitel protonu a „muž, který rozbil atom" (obrázek 2 [a]). Podobně jako jeho kolegové Rutherford už dlouhou dobu věděl, že je v atomových jádrech uloženo nesmírné množství energie; převládajícím názorem však bylo, že čerpat energii z tohoto zdroje není možné.

Dne 11. září 1933 se v Leicesteru konalo výroční setkání British Association for the Advancement of Science. Lord Rutherford měl proslov ve večerní části. Jak už to udělal předtím několikrát, odmítal jakékoli vyhlídky na atomovou energii: „Kdokoli hledá zdroj energie v transformaci atomů, jen mluví nesmysly." O Rutherfordově proslovu psali další den ráno v londýnských *The Times* (obrázek 2 [b]).

Maďarský fyzik Leo Szilard (obrázek 2 [c]), který před nedávnem utekl z nacistického Německa, zrovna přebýval v hotelu Imperial na londýnském Russellově



Obr. 2: (a) Jaderný fyzik lord Rutherford. (b) Výstřižky ze zprávy v *Timesech* z 12. září 1933 o Rutherfordově proslovu večer předchozího dne. (*Transformace prvků. Jaké jsou podle lorda Rutherforda výhledy na příštích 20–30 let? Je to ubohý a neefektivní způsob výroby energie. Kdokoli hledá zdroj energie v transformaci atomů, jen mluví nesmysly.*) (c) Leo Szilard, jaderný fyzik.

náměstí. Zprávu v *The Times* četl při snídani. S myšlenkou na to, co zrovna přečetl, si vyšel na procházku a přišel při tom na jadernou řetězovou reakci vyvolanou neutrony.[7] Problém osvobození jaderné energie přešel z nemožného do v podstatě vyřešeného během méně než 24 hodin. Další rok Szilard tajně podal patentovou přihlášku na jaderný reaktor. První patent na jadernou zbraň byl udělen ve Francii roku 1939.

Z této příhody plyne poučení, že je nerozum spoléhat na to, že lidský důmysl něco nedokáže, obzvlášť pak v situaci, kdy je v sázce budoucnost lidstva. V komunitě kolem AI se objevuje jistý druh popíračství, který jde dokonce tak daleko, že popírá možnost dosažení dlouhodobých cílů AI. Je to, jako kdyby řidič autobusu, který veze celé lidstvo, prohlásil: „Ano, jedu jednoznačně směrem k okraji propasti, ale důvěřujte mi – dřív než tam dojedeme, dojde mi palivo!"

Neříkám, že k úspěchu AI *nutně* dospějeme, a soudím, že je dost nepravděpodobné, že by se tak stalo v nejbližších letech. Zdá se nicméně prozíravé se na tuto možnost připravit. Pokud vše půjde dobře, ohlásí to zlatý věk lidstva, ale musíme se postavit čelem ke skutečnosti, že máme v plánu vytvořit entity, které jsou mnohem mocnější než lidé. Jak zajistíme, že nad námi skutečně nikdy nebudou mít moc?

Abychom získali aspoň představu, s jakým ohněm si zahráváme, uvažme jen fungování algoritmů na výběr obsahu na sociálních sítích. Nejsou nijak zvlášť inteligentní, ale dokážou působit na celý svět, protože přímo ovlivňují miliardy lidí. Tyto algoritmy jsou typicky navrženy tak, aby maximalizovaly *proklikávání*, tedy pravděpodobnost, že si uživatel na předkládanou položku klikne. Řešením je jednoduše uživateli předkládat položky, na které rád klikne, že? Nikoli. Řešením je změnit uživatelovy preference tak, aby se staly lépe předvídatelné. Předvídatelnějšího uživatele pak lze krmit položkami, u nichž je pravděpodobné, že na ně klikne, a tak generovat větší příjmy. Lidé s extrémnějšími politickými názory mají tendenci být předvídatelnější v tom, na jaké položky budou klikat. (Možná existuje kategorie článků, na které s velkou šancí kliknou zarytí centristé, ale není snadné si představit, o čem takové články budou.) Algoritmus se učí – podobně jako jakákoli racionální entita –, jak měnit stav prostředí, v němž pracuje (v tomto případě uživatelovo myšlení), aby tak maximalizoval přínos pro sebe sama.[8] Důsledky zahrnují revival fašismu, rozklad společenské smlouvy, na níž jsou postaveny demokracie po celém světě, a potenciálně i konec Evropské unie a NATO. To není špatné na několik řádků kódu, i pokud se mu dostalo pomocné ruky od některých lidí. A teď si představte, co by dokázal udělat *opravdu* inteligentní algoritmus.

CO SE POKAZILO?

Historie AI je poháněna jednoduchou mantrou: „Čím inteligentnější, tím lepší." Jsem přesvědčen o tom, že je to chyba – ne kvůli nějakému nejasnému pocitu, že nás někdo nahradí, ale kvůli způsobu, kterým chápeme inteligenci jako takovou.

Koncept inteligence je pro lidskou podstatu ústřední – proto se nazýváme *homo sapiens* neboli „člověk moudrý". Po více než dvou tisíciletích zkoumání sebe sama jsme dospěli k charakteristice inteligence, kterou lze shrnout takto:

> *Lidé jsou inteligentní do té míry, v níž se dá očekávat, že pomocí svého konání dosáhnou svých záměrů.*

Všechny ostatní charakteristiky inteligence – vnímání, myšlení, učení, vynalézání a tak – lze chápat skrze to, jak přispívají k naší schopnosti úspěšně konat. Od samých počátků AI byla inteligence strojů definována stejně:

> *Stroje jsou inteligentní do té míry, v níž se dá očekávat, že pomocí svého konání dosáhnou svých záměrů.*

Vzhledem k tomu že stroje na rozdíl od lidí nemají žádné vlastní záměry, zadáváme jim záměry, kterých mají dosáhnout, my. Jinými slovy, stavíme optimalizující stroje, nakrmíme je záměry a tradá, stroje jedou.

Tento obecný přístup neplatí jen pro AI. Opakuje se v technických i matematických základech naší společnosti. V oblasti teorie řízení, která se věnuje návrhu řídicích systémů pro cokoli od Jumbo Jetu po inzulinovou pumpu, je úkolem systému minimalizovat *nákladovou funkci*, která typicky měří nějakou odchylku od žádoucího chování. V oblasti ekonomie se navrhují mechanismy a politiky pro to, aby maximalizovaly *prospěch* jednotlivců, *blahobyt* skupin a *zisk* korporací.[9] V operačním výzkumu, který řeší komplexní logistické a výrobní problémy, maximalizuje řešení očekávanou *sumu odměn* během času. A konečně ve statistice jsou samoučící algoritmy navrženy tak, aby minimalizovaly očekávanou *funkci ztrát*, která definuje cenu důsledků chyb v předpovědi.

Toto obecné schéma – které nazýváme *standardním modelem* – je evidentně široce rozšířené a extrémně vlivné. Bohužel ale *nechceme stroje, které jsou inteligentní v tomto smyslu*.

Na slabou stránku standardního modelu poukázal roku 1960 Norbert Wiener, legendární profesor na MIT a jeden z nejlepších matematiků poloviny dvacátého století. Wiener tehdy viděl, jak se program Arthura Samuela učí hrát dámu mnohem lépe než jeho tvůrce. Tato zkušenost ho vedla k napsání prorockého, ale málo známého článku, „Some Moral and Technical Consequences

of Automation“ (Některé morální a technické důsledky automatizace).[10] Hlavní myšlenku v něm vyjadřuje takto:

> Pokud pro dosažení svých záměrů užijeme mechanizovaného činitele, do jehož operací se nemůžeme volně vměšovat, [...] měli bychom si být raději hodně jistí tím, že záměry vložené do stroje jsou záměry, které skutečně chceme.

„Záměr vložený do stroje“ je přesně tím účelem, který ve standardním modelu stroje optimalizují. Pokud vložíme špatný záměr do stroje, který je inteligentnější než my, dosáhne stroj svého cíle a my prohrajeme. Krize sociálních médií, kterou jsme popsali výše, je jen ochutnávkou – je jen výsledkem optimalizace špatného záměru v globálním měřítku za použití docela neinteligentních algoritmů. V kapitole 5 si vysvětlíme některé mnohem horší důsledky.

Nic z toho by nemělo být žádným velkým překvapením. Už tisíce let víme, jak to dopadá, když dostaneme přesně to, co jsme chtěli. V každém příběhu, kde se někomu plní tři přání, bývá třetí přání odčinit ta první dvě.

Celkově se zdá, že naše přibližování se nadlidské superinteligenci nelze zastavit, ale úspěch může znamenat vymazání lidské rasy. Nic ale není ztraceno. Musíme pochopit, kde jsme udělali chybu, a napravit ji.

DOKÁŽEME TO OPRAVIT?

Problém leží přímo v základní definici AI. Říkáme, že stroje jsou inteligentní do té míry, v níž se dá očekáva, že pomocí svého konání dosáhnou *svých* záměrů, ale nemáme žádný spolehlivý způsob, abychom zajistili, že *jejich* záměry jsou stejné jako *naše* záměry.

Co když namísto toho, abychom strojům dovolili sledovat *jejich* záměry, budeme trvat na tom, aby sledovaly *naše* záměry? Takový stroj, pokud by ho šlo navrhnout, by byl nejen *inteligentní*, ale také *prospěšný* lidem. Takže zkusme toto:

> *Stroje jsou **prospěšné** do té míry, v níž se dá očekávat, že pomocí **svého** konání dosáhnou **našich** záměrů.*

Tak jsme to nejspíš měli formulovat hned od začátku.

Problematické samozřejmě je, že naše záměry máme v sobě (všech osm miliard z nás, ve vší naší skvostné rozdílnosti), nikoli ve strojích. Je nicméně možné stavět stroje, které jsou prospěšné přesně v tomto smyslu. Takové stroje budou nevyhnutelně trpět nejistotou o našich záměrech – koneckonců si jimi nejsme jisti ani my sami –, ale ukazuje se, že toto je vlastnost, nikoli chyba (tedy že je to pozitivní, nikoli negativní znak). Nejistota ohledně cílů má za důsledek,

že stroje budou nutně podřízeny lidem: budou se ptát na dovolení, budou akceptovat opravy a dovolí, aby byly vypnuty.

Jakmile opustíme předpoklad, že by stroje měly mít určité záměry, budeme muset nahradit část základů umělé inteligence – základní definice toho, co se snažíme udělat. To také znamená přestavět velkou část nadstavby – nashromážděných myšlenek a metod, jak skutečně AI provozovat. Výsledkem bude nový vztah mezi lidmi a stroji, o němž doufám, že nám dovolí úspěšnou navigaci několika příštími dekádami.

KAPITOLA 2.

INTELIGENCE U LIDÍ A STROJŮ

Pokud dojdeme do slepé uličky, je dobrý nápad jít zpět po vlastních stopách a dopracovat se až k místu, kde jsme špatně zabočili. Uvedli jsme, že standardní model AI, v němž stroje optimalizují pevné záměry, jež jim předložili lidé, je slepou ulicí. Problémem není, že bychom mohli *selhat* v odvedení dobré práce na budování systémů AI; problém je naopak to, že bychom mohli *uspět*. Sama definice úspěchu v AI je chybná.

Vraťme se tedy po vlastních stopách celou cestu až k počátku. Snažme se pochopit, odkud se náš koncept inteligence vzal, a jak se stalo, že začal být používán na stroje. Pak budeme mít příležitost přijít s lepší definicí toho, co chápeme jako dobrý systém AI.

INTELIGENCE

Jak funguje vesmír? Jak začal život? Kde mám klíče? To jsou základní otázky, které si zaslouží naši pozornost. Ale kdo tyto otázky klade? Jak na ně odpovídáme? Jak může malé množství hmoty – pár dekagramů růžovošedého pudinku, které nazýváme mozkem – vnímat, chápat, předpovídat a měnit svět o nepředstavitelných rozměrech? A odsud už není daleko k tomu, aby se myšlení obrátilo ke zkoumání sebe sama.

Po tisíce let jsme se snažili porozumět, jak funguje naše myšlení. Na počátku tomu tak bylo mimo jiné proto, že nás k tomu vedla zvědavost, sebeovládání, přesvědčení a také docela pragmatický cíl analýzy matematické argumentace. Přesto ale každý krok směrem k vysvětlení, jak myšlení pracuje, znamená současně také krok směrem ke vzniku schopnosti myšlení v nějakém artefaktu – to znamená krok směrem k umělé inteligenci.

Než budeme moci porozumět tomu, jak inteligenci vytvořit, je užitečné porozumět tomu, co to vlastně je. Odpověď nenajdeme v testech IQ, a dokonce ani v Turingových testech, ale v jednoduchém vztahu mezi tím, co vnímáme, co chceme a co děláme. Zhruba řečeno – entita je inteligentní do takové míry, v níž má šanci dosáhnout toho, co chce, v závislosti na tom, co vnímá.

VÝVOJOVÉ POČÁTKY

Vezměme si jednoduchou bakterii, třeba *Escherichii coli*. Je vybavena asi půltuctem bičíků – dlouhých chapadel podobných vlasům, které kolem své základny rotují po nebo proti směru hodinových ručiček. (Motor tohoto otáčení je úžasná věc sama o sobě, ale to je jiný příběh.) Když *E. coli* proplouvá ve svém tekutém domově – ve vašich střevech –, střídá rotaci svých bičíků po směru hodinových ručiček, což způsobuje, že se „válí" na místě, a proti směru hodinových ručiček, což vede k tomu, že se bičíky kolem sebe ovinou do jakéhosi pohonného mechanismu, díky němuž bakterie plave v přímém směru. To znamená, že *E. coli* vykonává něco jako nahodilou procházku – plav, válej se, plav, válej se –, při níž dokáže hledat a konzumovat glukózu a vyhnout se tomu, aby se nehnula z místa a zemřela hladem.

Pokud by to bylo všechno, neřekli bychom o *E. coli*, že je nějak zvlášť inteligentní, protože její konání nijak nezávisí na prostředí, v němž se nachází. Nedělala by žádná rozhodnutí, pouze by realizovala pevně dané chování, které jí evoluce vystavěla v genech. Ale to není celé. Pokud *E. coli* ucítí zvýšenou koncentraci glukózy, plave déle a válí se méně, a dělá opak, když cítí, že se koncentrace glukózy snižuje. Takže tím, co dělá (plavba směrem ke glukóze), má šanci dosáhnout toho, co chce (předpokládejme, že jde o více glukózy), na základě toho, co vnímá (změna koncentrace glukózy).

Teď si asi pomyslíte: „Ale evoluce to vložila do jejích genů! Jak ji to může dělat inteligentní?" Tohle je nebezpečný směr uvažování, protože evoluce také do vašich genů vložila základní plány vašeho mozku a nejspíš na základě toho nebudete chtít popřít svou vlastní inteligenci. Jde o to, že to, co evoluce vložila do genů *E. coli* (stejně jako do vašich), je mechanismus, pomocí něhož se chování bakterie mění podle toho, co vnímá v prostředí kolem sebe. Evoluce předem neví, kde se bude glukóza nacházet nebo kde leží vaše klíče, a tak vložit do organismu schopnosti takové věci hledat je to nejlepší, co se dá dělat.

E. coli samozřejmě není žádný intelektuální gigant. Pokud víme, nepamatuje si, kde se nacházela, takže pokud se přesune z A do B a nenajde žádnou glukózu, může se stejně pravděpodobně vrátit zpět do A. Pokud sestavíme prostředí, v němž každý přitažlivý gradient glukózy vede k místu, kde je fenol (což je pro *E. coli* jed), bakterie bude tyto gradienty sledovat. Nikdy se nic nenaučí. Nemá žádný mozek, probíhá v ní jen několik jednoduchých chemických reakcí, které určují výsledek.

Velký krok kupředu se objevil s akčními potenciály, což je forma elektrických signálů, které se nejdříve rozvinuly v jednobuněčných organismech zhruba před miliardou let. Pozdější vícebuněčné organismy si vyvinuly specializované buňky zvané *neurony*, které používají elektrický akční potenciál k rychlému přenosu signálů – až 120 m za sekundu neboli 430 km za hodinu – uvnitř organismu.