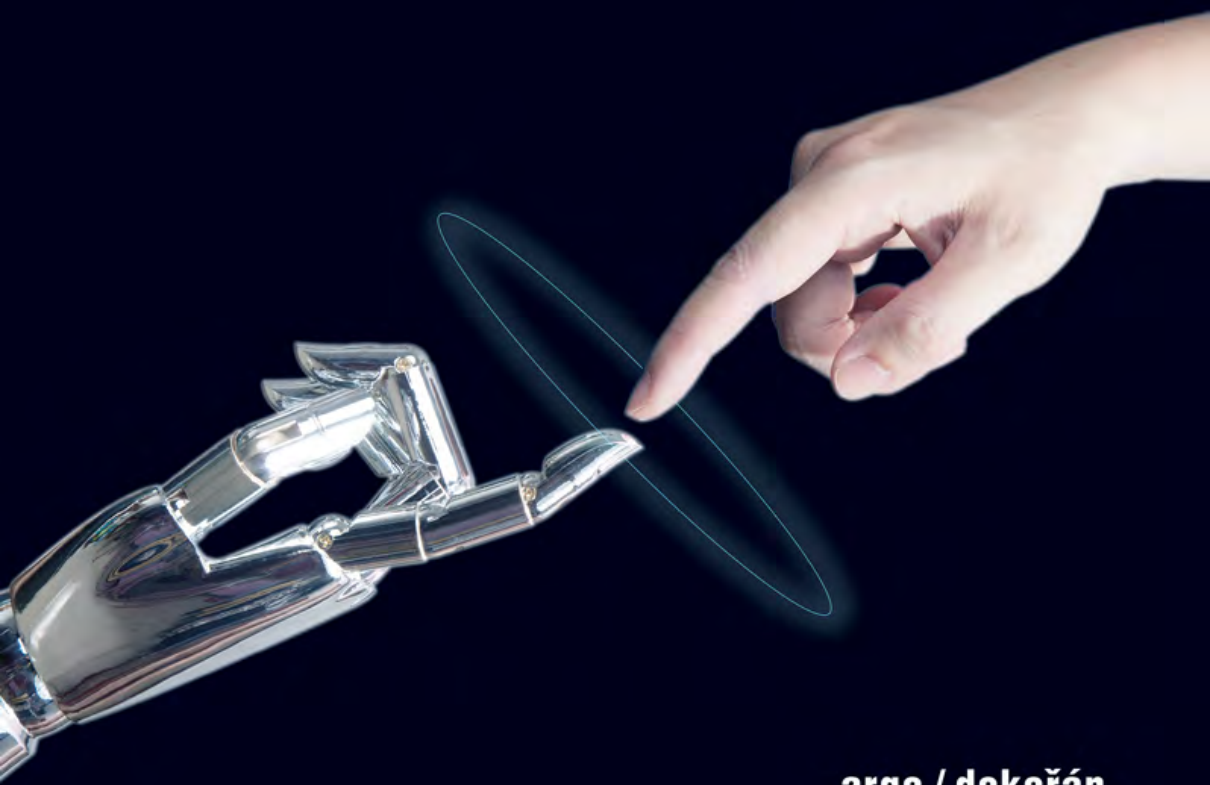


Stuart Russell

# JAKO ČLOVĚK

Umělá inteligence a problém jejího ovládní



argo / dokořán



Stuart Russell

# JAKO ČLOVĚK

**Umělá inteligence a problém jejího ovládní**

**ARGO / DOKOŘÁN**

Stuart Russell

# Jako člověk

**Umělá inteligence a problém jejího ovládní**

Copyright © 2019 by Stuart Russell. All rights reserved.

Translation © Jiří Zlatuška, 2021

Všechna práva vyhrazena. Žádná část této publikace nesmí být rozmnožována a rozšiřována jakýmkoli způsobem bez předchozího písemného svolení nakladatele.

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* 7

KAPITOLA 1.	<b>Pokud uspějeme</b>	9
	Jak jsme sem došli?	11
	Co se stane dál?	13
	Co se pokazilo?	15
	Dokážeme to opravit?	16
KAPITOLA 2.	<b>Intelligence u lidí a strojí</b>	19
	Intelligence	19
	Počítače	32
	Inteligentní počítače	38
KAPITOLA 3.	<b>Jaký vývoj AI čeká v budoucnu?</b>	55
	Blízká budoucnost	55
	Kdy přijde superinteligentní AI?	65
	Konceptuální průlomů před námi	66
	Představy o superinteligentním stroji	77
	Meze superintelligence	79
	Jak bude AI prospěšná pro člověka?	81
KAPITOLA 4.	<b>Zneužití AI</b>	85
	Stálý dohled, přesvědčování a ovládání	85
	Smrtící autonomní zbraně	90
	Likvidace práce v podobě, jak ji známe	92
	Uchvácení dalších lidských rolí	100
KAPITOLA 5.	<b>Příliš inteligentní AI</b>	107
	Problém gorily	107
	Problém krále Midase	110
	Strach a chamtivost: instrumentální cíle	113
	Intelligenční exploze	114
KAPITOLA 6.	<b>Nikterak velkolepá debata o AI</b>	117
	Popření	118
	Odchylování	123
	Tribalismus	126
	Nemůžeme prostě...	127
	Znovunastartovaná debata	134

KAPITOLA 7.	<b>AI: odlišný přístup</b>	135
	Principy pro prospěšné stroje	136
	Důvody k optimismu	140
	Důvod k opatrnosti	142
KAPITOLA 8.	<b>Dokazatelně přínosná AI</b>	145
	Matematické záruky	145
	Učit se preference z chování	149
	Asistenční hry	151
	Žádosti a pokyny	159
	Zadrátování hlavy	161
	Rekurzivní sebezdokonalování	163
KAPITOLA 9.	<b>Komplikace: my</b>	165
	Různí lidé	165
	Mnoho lidí	166
	Hodní, protivní a závistiví lidé	176
	Hloupí, emocionální lidé	179
	Mají lidé opravdu preference?	181
KAPITOLA 10.	<b>Vyřešený problém?</b>	189
	Prospěšné stroje	189
	Ovládání AI	191
	Zneužití	193
	Ochablost a lidská autonomie	194
PŘÍLOHA A.	<b>Hledání řešení</b>	197
	Rezignace na racionální rozhodování	198
	Pohled do budoucna	201
PŘÍLOHA B.	<b>Znalosti a logika</b>	205
	Výroková logika	205
	Logika prvního řádu	207
PŘÍLOHA C.	<b>Nejistota a pravděpodobnost</b>	209
PŘÍLOHA D.	<b>Učení se ze zkušeností</b>	218
	<i>Poděkování</i>	226
	<i>Poznámky</i>	227
	<i>Zdroje ilustrací</i>	259
	<i>Rejstřík</i>	261

# 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 všeprostopupujícím aspektem současnosti, ale proto, že jde o dominantní technologii budoucnosti. Světové mocnosti si této skutečnosti začínají všimnout 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

Knih 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 ovládnutí: 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.





# 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 profesori* 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.

Zvlášt 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.<sup>1</sup> 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 pojmut 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 schouliil. (Další upozornění z *katedry shod okolnosti*?) 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 Rochesterera, 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<sup>2</sup> a program hrající dámu od Arthura Samuela, který se naučil porážet

svého tvůrce.<sup>3</sup> 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.“<sup>4</sup> 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<sup>5</sup> na Imperial College, napsal jsem program na hraní šachů – štos dřevných 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 zasetá 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.<sup>6</sup>

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?

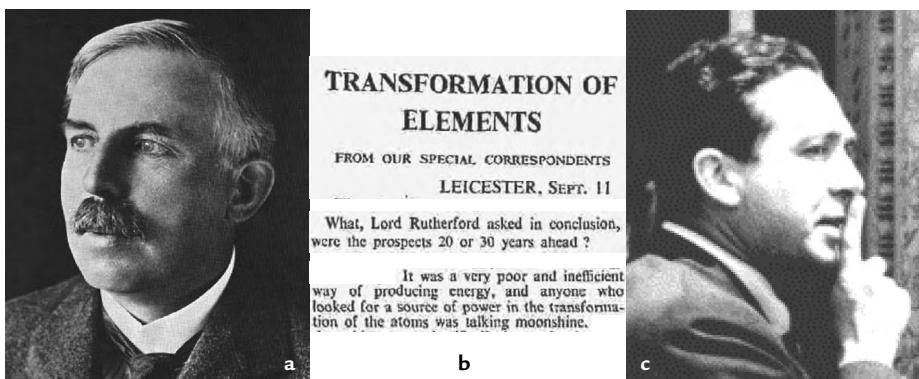
Znamená takto rychlé tempo pokroku, že budeme za chvíli předstíženi stroji? Ne. Je zde několik průlomových událostí, které se musí stát, než budeme mít cokoli, co by připomínalo nadlidskou inteligenci.

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.<sup>7</sup> 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 uživatelské 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 uživatelské myšlení), aby tak maximalizoval přínos pro sebe sama.<sup>8</sup> 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 navrhuje mechanismy a politiky pro to, aby maximalizovaly *prospěch* jednotlivců, *blahobyt* skupin a *zisk* korporací.<sup>9</sup> 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 pro-rockého, ale málo známého článku, „Some Moral and Technical Consequences

of Automation“ (Některé morální a technické důsledky automatizace).<sup>10</sup> 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 jisti 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 očekávat, 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 vši 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.



# INTEIGENCE 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.

## INTEIGENCE

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 bičičků – 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* proplová ve svém tekutém domově – ve vašich střevech –, střídá rotaci svých bičičků po směru hodinových ručiček, což způsobuje, že se „válí“ na místě, a proti směru hodinových ručiček, což vede k tomu, že se bičičky 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 jejich 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.

Tato spojení mezi neurony se nazývají *synapse*. Síla synaptických spojů určuje, jaké množství se z elektrického nabuzení přenáší z jednoho neuronu na druhý. Změnou síly synaptických spojení se zvířata učí.<sup>1</sup> Učení poskytuje ohromnou evoluční výhodu, protože se zvířata mohou adaptovat na velké rozpětí různých okolností. Učení také zrychluje tempo evoluce samé.

Zpočátku byly neurony uspořádány do nervových sítí, které jsou rozprostřeny po celém organismu a slouží ke koordinaci činností jako vyživování a trávení nebo k synchronizaci smršťování svalových buněk. Elegantní pohyb medúz je výsledkem práce nervových sítí. Medúzy žádný mozek nemají.

Mozky vznikly později, spolu s komplexními smyslovými orgány, jako jsou oči a uši. Několik stovek milionů let poté, co se objevily medúzy se svými nervovými sítěmi, jsme přišli my lidé s velkými mozky – se stovkami miliard ( $10^{11}$ ) neuronů a biliardou ( $10^{15}$ ) synapsí. I když jsou pomalé ve srovnání s elektronickými obvody, „časovací cyklus“ několika málo milisekund na změnu stavu je ve srovnání s většinou biologických procesů rychlý. Lidský mozek jeho majitelé často popisují jako „nejsložitější objekt ve vesmíru“, což sice asi není pravda, ale slouží to jako dobrá výmluva pro fakt, že stále ještě moc nechápeme, jak skutečně pracuje. Víme toho sice mnoho o biochemii neuronů a synapsí a o anatomických strukturách mozku, ale na nervovou implementaci *kognitivní* úrovně – učení, vědění, paměti, uvažování, plánování, rozhodování a podobně – panuje pořád ještě řada různých názorů.<sup>2</sup> (To se možná změní, až toho pochopíme víc o AI nebo až vyvineme přesnější nástroje na měření mozkové aktivity.) Když tedy čteme v médiích, že taková a taková technika v AI „pracuje úplně stejně jako lidský mozek“, bude to buď jen něčí odhad, nebo dokonce čirá smyšlenka.

O oblasti *vědomí* ve skutečnosti nevíme nic, takže také nic říkat nebudu. Nikdo v AI nepracuje na tom, aby udělal stroje s vědomím, a nikdo ani netuší, odkud začít, navíc žádné chování nepotřebuje vědomí jako východisko. Předpokládejme, že vám ukázu nějaký program a zeptám se: „Představuje hrozbu pro lidstvo?“ Vy jeho kód zanalyzujete a vskutku, při jeho běhu tento program vytvoří a provede plán, jehož výsledkem bude zničení každého člověka, který se mu postaví, podobně jako šachový program vytvoří a provede plán, jehož výsledkem bude porážka každého člověka, s nímž hraje. Předpokládejme nyní, že vám dále řeknu, že jakmile se ten kód bude provádět, vytvoří také nějakou formu strojového vědomí. Změní to vaši předpověď? Ani trochu. Nebude to znamenat *naprosto žádný rozdíl*.<sup>3</sup> Vaše předpověď o jeho chování bude přesně stejná, protože je tato předpověď založena na kódu. Všechny ty hollywoodské zápletky o strojích, které nějak záhadně získaly vědomí a začaly nenávidět lidi, ve skutečnosti míří vedle: o co jde, jsou schopnosti, nikoli vědomí.

Existuje jeden důležitý kognitivní aspekt mozku, kterému rozumět *opravdu* začínáme – konkrétně jde o *systém odměňování*. Jde o vnitřní signální systém,

zprostředkovávaný dopaminem, jenž spojuje pozitivní a negativní stimuly chování. Jeho fungování bylo objeveno švédským neurovědcem Nilsem-Åkem Hillarpem a jeho spolupracovníky na konci 50. let 20. století. Způsobuje, že vyhledáváme pozitivní stimulace, jako třeba sladká jídla, které zvyšují dopaminovou hladinu, a vede nás k tomu, že se vyhýbáme negativní stimulaci jako hladu nebo bolesti, která dopaminovou hladinu snižuje. V jistém smyslu je to podobné mechanismu na vyhledávání glukózy u *E. coli*, ale mnohem komplexnější. Součástí tohoto systému je metoda učení, takže se naše chování stává během času při získávání odměny úspěšnějším. Umožňuje také uspokojení odkládat, takže se naučíme přát si věci jako peníze, které poskytují odměnu až později, nikoli okamžitě. Jedním důvodem, proč systému získávání odměn v mozku rozumíme, je to, že připomíná metodu *zpětnovazebního učení* vyvinutou v AI, pro kterou máme velmi solidní teorii.<sup>4</sup>

Z evolučního hlediska lze přemýšlet o systému odměn v mozku stejně jako o mechanismu vyhledávání glukózy u *E. coli*, tedy jako o způsobu zlepšování evolučních způsobilostí. Organismy, které jsou při vyhledávání odměny úspěšnější - tedy naleznou dobré jídlo, vyhnou se bolesti, zapojí se do sexuálních aktivit a podobně -, budou mít větší šanci, že rozšíří své geny. Pro jednotlivý organismus je nesmírně obtížné rozhodnout, které činnosti vedou v dlouhodobém výhledu k nejúspěšnějšímu šíření vlastních genů, takže nám to evoluce zjednodušila tím, že nám poskytuje zabudované směrníky.

Tyto směrníky ovšem nejsou dokonalé. Existují způsoby, jak získat odměnu, které nejspíš snižují pravděpodobnost toho, že se geny jejich nositele rozšíří. Tak například užívání drog, pití velkého množství slazených bublinkových nápojů nebo hraní videoher osmnáct hodin denně bude nejspíš kontraproduktivní z hlediska toho, co je ve hře při reprodukci. Pokud bychom navíc dostali možnost přímého elektrického přístupu do svého systému odměn, pravděpodobně bychom sebe samé stimulovali bez přestání, dokud bychom nezemřeli.<sup>5</sup>

Nesrovnalosti mezi systémem signálů odměňování a evoluční způsobilostí nemají dopad jen na jednotlivá izolovaná individua. Na ostrůvku kousek od pobřeží Panamy žije tříprstý trpasličí lenochod, který je zřejmě závislý na látce podobné Valiu (diazepamu), kterou obsahuje jeho potrava z červených listů mangrovníku, a hrozí mu vyhynutí.<sup>6</sup> Vypadá to tak, že může zmizet celý živočišný druh, pokud si najde niku, kde může uspokojovat svůj systém odměn špatným způsobem.

Kromě těchto výjimek ovšem to, že se lze naučit maximalizovat své odměny v přírodních prostředích, obvykle zlepšuje šance organismů šířit své geny a přezít změny životního prostředí.

## EVOLUČNÍ URYCHLOVAČ

Učení je dobré pro víc věcí než jen přežívání a prospívání. Také *urychluje evoluci*. Jak je to možné? Učení přece nemění DNA organismu, celá evoluce se týká změn DNA za více generací. Se vztahem mezi učením a evolucí přišel v roce 1896 americký psycholog James Baldwin<sup>7</sup> a nezávisle na něm i britský etolog Conwy Lloyd Morgan<sup>8</sup>, ale jejich myšlenky v té době nebyly obecně akceptovány.

Baldwinův efekt, jak je dnes znám, lze chápat pomocí představy, že evoluce má možnost volby mezi vytvořením *instinktivního* organismu, jehož každá odezva je předem fixována, a vytvořením *adaptivního* organismu, jenž se učí, jaké kroky udělat. A teď pro ilustraci předpokládejme, že optimální instinktivní organismus lze zakódovat šestimístným číslem, řekněme 472116, zatímco v případě adaptivního organismu evoluce specifikuje pouze 472\*\*\* a je na organismu samém, aby vyplnil poslední tři číslice tím, že se během celého života bude učit. Zjevně má evoluce práci výrazně usnadněnu, pokud se má starat jen o výběr prvních tří číslic; adaptivní organismus během učení, jaké jsou poslední tři číslice, provádí v jednom životě totéž, co by evoluce musela dělat po mnoho generací. Pokud adaptivní organismy během učení přežijí, vypadá to, že schopnost učení představuje evoluční zkratku. Z počítačových simulací se zdá, že Baldwinův efekt je reálný.<sup>9</sup> Působení kultury tento proces také zrychluje, protože organizovaná civilizace chrání jednotlivý organismus během doby, kdy se učí, a předává mu informace, které by se tento jedinec jinak musel naučit sám.

Příběh Baldwinova efektu je fascinující, ale není úplný: předpokládá, že učení a evoluce nutně směřují stejným směrem, tedy že ať interní signál zpětné vazby definuje v organismu směr učení jakkoli, bude to v dokonalé shodě s evoluční způsobilostí. Jak jsme ale viděli na příkladu tříprstého trpasličího lenochoda, nemusí to být vždy pravda. V nejlepším případě poskytnou zabudované mechanismy učení se pouze hrubý náznak dlouhodobých dopadů jakéhokoli konkrétního kroku na evoluční způsobilost. Navíc je třeba se ptát: „Jak se tam systém odměn vůbec dostal?“ Odpověď samozřejmě zní, že se tak stalo evolučním procesem, který internalizoval mechanismus zpětné vazby, jenž je aspoň do nějaké míry souvztažen s evoluční způsobilostí.<sup>10</sup> Učící mechanismus, který by působil tak, že by se organismy vzdalovaly od potenciálních partnerů a utíkaly směrem k predátorům, by zjevně nevydržel moc dlouho.

Baldwinovu efektu děkujeme za to, že jsou neurony se svými schopnostmi pro učení se a řešení problémů ve zvířecí říši natolik rozšířené. Současně je důležité chápat, že se evoluce ve skutečnosti nestará o to, zda máme mozek nebo zda přemýšlíme o zajímavých myšlenkách. Evoluce nás považuje jen za aktéra (*agenta*), to znamená za něco, co koná. Takové zásadní intelektuální charakteristiky jako logické uvažování, cílené plánování, moudrost, zdravý rozum, představivost a tvůrčí schopnosti mohou být pro vytvoření inteligentního aktéra podstatné,

ale také nemusí. Jedním důvodem, proč je umělá inteligence tak úchvatná, je její schopnost poskytnout cestu k pochopení těchto záležitostí: můžeme dospět k poznání, jak tyto intelektuální charakteristiky vůbec umožňují inteligentní život, i proč je nemožné bez nich vyprodukovat skutečně inteligentní chování.

### RACIONALITA PRO JEDNOHO

Od nejranějších počátků starověké řecké filozofie byl koncept inteligence navázán na schopnosti vnímat, uvažovat a *úspěšně* konat.<sup>11</sup> Během staletí se tento koncept rozšířil co do možných aplikací a jeho definice se zpřesnila.

Aristoteles (spolu s dalšími mysliteli) se věnoval studiu pojmu úspěšného uvažování – metodám logické dedukce, která za předpokladu pravdivosti premis vede k pravdivým závěrům. Zabýval se také zkoumáním procesu rozhodování, jak konat – někdy označovaným za *praktické uvažování* –, a přišel s představou, že obsahuje předpoklad, že jisté druhy konání dokážou dosáhnout zamýšleného cíle:

Neuvažujeme tedy o účelech a cílech, nýbrž o prostředcích. Neboť ani lékař neuvažuje, bude-li léčit, ani řečník, bude-li přemlouvav [...]. Ale když si stanovili cíl, hledí k tomu, jak a čím ho dosíci, a jeví-li se, že ho lze dosíci více prostředky, přihlížejí k tomu, kterým prostředkem lze ho dosáhnout nejsnáze a nejkrásněji. Možno-li však uskutečnit jej jen jedním prostředkem, uvažují, *jak* jest možno uskutečnit jej *tímto* a čím zase onen, až přijdou k příčině, která v hledání jest poslední; [...] to, co při rozboru jest poslední, jest při uskutečňování první. A narazí-li na nemožnost, ustávají, například je-li potřeba peněz a není možno je opatřit; ale jeví-li se to možným, pokoušejí se jednat.<sup>12</sup>

Dalo by se říct, že tato pasáž nastavila směr dalších zhruba dvou tisíc let západních úvah o racionalitě. Říká, že „konec“ – to, co člověk chce – je pevný a daný; a říká, že racionální konání je takové, které podle logických dedukcí napříč posloupností činů „jednoduše a nejlépe“ produkuje tento konec.

Aristotelův pohled vypadá rozumně, ale není to úplný průvodce racionálním chováním. Pomíjí konkrétně záležitost nejistoty. Ve skutečném světě má realita tendenci se do dění vkládat a jen u mála kroků nebo jejich posloupností je skutečně zaručeno, že dosáhnou zamýšleného konce. Tak například teď, když píšu tuto větu, je v Paříži deštivá neděle a v úterý ve 14:30 odlétám letadlem z letiště Charlese de Gaulla, zhruba tři čtvrtě hodiny od mého domu. Plánuji na letišti odjet kolem 11:30, takže budu mít dostatek času, ale taky to znamená, že na letišti budu nejméně hodinu sedět v odletové oblasti. Stihnu ten let *jistě*? Nikoli. Můžu se dostat do dopravní zácpy, řidiči taxíků mohou stávkovat, taxi, kterým pojedou, se může rozbít nebo mohou jeho řidiče zavřít



po rychlé honičce ulicemi a podobně. Mohl bych místo toho na letišti odjet už v pondělí, o celý den napřed. To by výrazně snížilo šanci zmeškat let, ale vyhlídka na noc v odletovém salonku není nijak přitažlivá. Jinými slovy, můj plán zahrnuje *substituční vztah* mezi jistotou úspěchu a cenou za jeho zajištění. Následující záměr koupě domu obsahuje vztah něco za něco v podobném smyslu: kupte si tiket do loterie, vyhraje milion dolarů a pak si kupte dům. Takový plán „jednoduše a nejlépe“ produkuje kýžený konec, ale je krajně nepravděpodobné, že by mohl uspět. Rozdíl mezi takto ztřeštěným plánem na nákup domu a mým střízlivým plánem na jízdu na letišti je ovšem jen záležitostí míry. V obou případech jde o hazard, ale jeden z nich vypadá racionálněji než ten druhý.

Ukazuje se, že hazardování hrálo ústřední roli v Aristotelových úvahách, jak se vypořádat s neurčitostí. Po roce 1560 vyvinul italský matematik Gerolamo Cardano první matematicky přesnou teorii pravděpodobnosti - jako hlavní příklad užíval hru v kostky. (Naneštěstí byla tato práce publikována až v roce 1663.<sup>13</sup>) V sedmnáctém století francouzští myslitelé včetně Antoina Arnaulda a Blaise Pascala začali - z nepochybných matematických důvodů - studovat otázku racionálních rozhodnutí v hazardu.<sup>14</sup> Vezměme si dvě sázky:

A: 20procentní šance vyhrát 10 dolarů

B: 5procentní šance vyhrát 100 dolarů

Návrh, s nímž matematici přišli, je nejspíš stejný, jak bychom postupovali my sami: srovnáme *očekávané hodnoty* sázek, tedy průměrnou sumu, kterou můžeme očekávat, že z každé sázky získáme. Pro sázku A je očekávaná hodnota 20 procent z 10, neboli 2 dolary. Pro sázku B je očekávaná hodnota 5 procent ze 100, neboli 5 dolarů. Sázka B je tak podle této teorie lepší. Tahle teorie dává smysl, protože pokud opakujeme stále stejné sázky, pak sázející, který se podle toho chová, skončí s větším množstvím peněz než ten, kdo se tak nechová.

V osmnáctém století švýcarský matematik Daniel Bernoulli postřehl, že toto pravidlo moc nefunguje pro větší sumy.<sup>15</sup> Vezměme si následující dvě sázky:

A: 100procentní šance získat 10 000 000 dolarů

(očekávaná hodnota 10 milionů dolarů)

B: 1procentní šance získat 1 000 000 100 dolarů

(očekávaná hodnota 10 milionů a 1 dolar)

Většina čtenářů této knihy by (stejně jako její autor) preferovala sázku A před sázkou B, i když pravidlo očekávané hodnoty říká opak! Bernoulli předpokládal, že se sázky vyhodnocují nikoli podle očekávané peněžní hodnoty, nýbrž podle

jejich očekávaného *užitku*. Užitek – vlastnost být užitečným nebo prospěšným nějaké osobě – byl, jak předkládal, vnitřní subjektivní kvantitou se vztahem k peněžní hodnotě, ale různou od ní. Užitek zejména vykazuje *klesající výnosy z rostoucích částek*. To znamená, že užitek nějaké sumy peněz není striktně úměrný této sumě, ale od nějaké hodnoty roste pomaleji než suma sama. Tak třeba užitek z vlastnictví 1 000 000 100 dolarů je mnohem menší než stonásobek užítku vlastnictví 10 000 001 dolaru. Kolikrát menší? Dával jsem tuto otázku postgraduálním studentům ve svém předmětu a jejich odpovědi byly kolem 50, tedy že sázka B by měla mít očekávanou hodnotu 500 milionů dolarů, aby dosáhla lákavosti sázky A. Řeknu to ještě jednou: sázka B by měla očekávanou hodnotu padesátkrát větší než sázka A, aby obě sázky měly stejný užitek.

Bernoulliho zavedení užítka – neviditelné vlastnosti – pro vysvětlení lidského chování pomocí matematické teorie bylo na svou dobu naprosto nečekanou myšlenkou. A byla ještě pozoruhodnější kvůli tomu, že na rozdíl od očekávaných hodnot nejsou očekávané užítky různých druhů sázek nebo cen přímo měřitelné; užitek se musí *odvodit z preferencí* každého jednotlivce. Musela uplynout ještě dvě staletí, než byly plně propracovány důsledky této teze a než došlo širokého přijetí mezi statistiky a ekonomy.

V polovině 20. století to byli John von Neumann (velký matematik, po němž je pojmenována standardní „von Neumannova architektura“ počítačů<sup>16</sup>) a Oscar Morgenstern, kdo publikovali *axiomatickou* bázi pro teorii užítka.<sup>17</sup> Znamená to následující: dokud preference zastávané nějakým jednotlivcem splňují nějaké základní axiomy vhodné pro každého racionálního aktéra, pak volby, které jednotlivec dělá, lze nutně popsat jako maximalizaci očekávané hodnoty funkce užítka. Krátce řečeno, *racionální aktér jedná tak, aby maximalizoval očekávaný užitek*.

Důležitost tohoto závěru je obrovská. Umělá inteligence se v mnohém smyslu zabývá vypracováváním detailů toho, jak vytvořit racionální stroje.

Podívejme se trochu podrobněji na axiomy, jejichž splnění u každé racionální entity očekáváme. Zde je jeden, který se nazývá *tranzitivita*: pokud upřednostňujeme cenu A před cenou B a máme na vybranou mezi loteriami, kde jsou A a B jedinými možnými výstupy, pak upřednostňujeme loterii, kde má A vyšší šanci než B. To zní rozumně! (Pokud máte raději salámovou pizzu než margaritu a margaritu než ananasovou, pak zní rozumně, že si zvolíte salámovou místo ananasovou.) Jiný se nazývá *monotonicitá*: pokud upřednostňujete cenu A před B a můžete si vybrat z loterií, kde jsou A a B jedinými možnými výstupy, upřednostníte loterii s nejvyšší pravděpodobností výhry A, nikoli B. To je zase zcela rozumné.

Preferenze se netýkají jen pizzy nebo loterie s peněžními výhrami. Mohou se týkat úplně čehokoli; zejména se pak mohou týkat čehokoli z našeho života

nebo ze životů jiných lidí. Pokud se zabýváme preferencemi, které zahrnují posloupnosti událostí v čase, existuje další často činěný předpoklad, nazývaný *stacionárnost*: pokud dvě různé budoucnosti A a B začínají stejnou událostí a preferujeme A před B, pak preferujeme A před B i poté, co se tato událost stala. Zní to rozumně, ale má to to překvapivě silný důsledek: užitek libovolné posloupnosti událostí je součet odměn spojených s každou z událostí (s možným diskontováním v čase, něčím jako subjektivní úrokovou mírou).<sup>18</sup> I když je tento předpoklad „užitku jako součtu odměn“ historicky rozšířen – přinejmenším od osmnáctého století, od „hedonického kalkulu“ zakladatele utilitarismu Jeremyho Benthamy –, není předpoklad stacionárnosti, na němž je založen, nezbytnou vlastností racionálních aktérů. Stacionárnost také vylučuje možnost, že se něčí preference mohou měnit v čase, zatímco naše zkušenost svědčí o opaku.

Bez ohledu na racionalitu axiomů a důležitost závěrů, které z nich plynou, byla teorie užitku neustále bombardována kritickými námitkami od doby, kdy se s ní poprvé seznámila veřejnost. Někteří ji zavrhují, protože údajně vše redukuje na peníze a sobeckost. (Této teorii se někteří francouzští autoři vysmívali jako „americké“, i když má svůj původ ve Francii.) Ve skutečnosti je ale zcela racionální chtít žít život v odříkání a chtít jen omezovat utrpení ostatních. Altruismus jednoduše znamená přičítat při vyhodnocování jakékoli předložené budoucnosti podstatnou váhu tomu, jak se vede druhým.

Jiná sada námitek souvisí s obtížemi při zjišťování potřebných pravděpodobností a hodnot užitku a jejich vzájemném násobení při výpočtu očekávaných užiteků. Tyto námitky jednoduše pletou dvě různé věci: volbu racionálního konání a jeho výběr pomocí *výpočtu očekávaných užiteků*. Pokud se třeba snažíme píchnout prstem do oka, zavřou se naše oční víčka bez jakýchkoli propočtů; je to racionální, ale nejsou do toho zataženy žádné výpočty očekávaného užitku. Nebo předpokládejme, že jedeme na kole z kopce bez brzd a máme na výběr mezi nárazem do jedné betonové zdi rychlostí patnáct kilometrů v hodině nebo do jiné o něco níž rychlostí třicet kilometrů v hodině; kterou z nich si vybereme? Pokud jste si vybrali patnáct kilometrů v hodině, pak gratuluji. Počítali jste si očekávané hodnoty užitku? Nejspíš nikoli. Ale upřednostnění patnácti kilometrů za hodinu je pořád racionální.<sup>19</sup> Plyne to ze dvou základních předpokladů: zaprvé dáváme přednost méně vážným zraněním před vážnějšími a zadruhé pro jakoukoli úroveň zranění platí, že zvýšení rychlosti srážky zvyšuje šanci překročení této úrovně. Z těchto dvou předpokladů plyne matematickou úvahou – bez užití jakýchkoli čísel –, že narazit v rychlosti 15 kilometrů v hodině má vyšší očekávaný užitek než srážka v rychlosti 30 kilometrů v hodině.<sup>20</sup> Celkově nemusí maximalizace očekávaného užitku vyžadovat žádné výpočty očekávání nebo užiteků. Jde o čistě *externí* popis racionální entity.