

NEW YORK TIMES BESTSELLER ★★ 2014

NICK BOSTROM



SUPER INTELI GENCE

AŽ BUDOU STROJE CHYTŘEJŠÍ NEŽ LIDÉ





Nick Bostrom

Superintelligence

Až budou stroje chytřejší než lidé

PROSTOR

NICK BOSTROM SUPER INTELI GENCE

AŽ BUDOU STROJE CHYTŘEJŠÍ NEŽ LIDÉ

přeložil Jan Petříček



PROSTOR | PRAHA | 2018

SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES, FIRST EDITION
was originally published in English in 2014. This translation is published
by arrangement with Oxford University Press.

Copyright © Nick Bostrom, 2014
All rights reserved

Czech edition © PROSTOR, 2018
Translation © Jan Petříček, 2017
Cover picture © shutterstock/isifa, 2018

ISBN 978-80-7260-389-3

OBSAH

<i>Nedokončená bajka o vrbcích</i>	11
--	----

Předmluva	13
----------------------------	----

kapitola 1

Dosavadní vývoj a dnešní schopnosti	17
Režimy růstu a velké dějiny	17
Nadějné vyhlídky	20
Období naděje a zoufalství	23
Současný stav oboru	33
Názory na budoucnost strojové inteligence	42

kapitola 2

Cesty k superinteligenci	47
Umělá inteligence	48
Úplná emulace mozku	58
Biologická kognice	68
Rozhraní mozek-počítač	80
Sítě a organizace	86
Shrnutí	88

kapitola 3

Formy superinteligence	91
Rychlostní superinteligence	92
Kolektivní superinteligence	93
Kvalitativní superinteligence	97
Přímý a nepřímý dosah	99
Výhody digitální inteligence	101

kapitola 4

Kinetika intelligenční exploze	106
Doba a rychlost vzestupu	106
Systémový odpor	112
<i>Cesty nezahrnující strojovou inteligenci</i>	112
<i>Cesta emulace a UI</i>	114
Optimalizační výkon a explozivnost	123

kapitola 5

Rozhodující strategická výhoda	129
Získá vedoucí projekt rozhodující strategickou výhodu?	130
Velikost úspěšného projektu	136
<i>Monitorování</i>	137
<i>Mezinárodní spolupráce</i>	140
Od rozhodující strategické výhody ke globální samovládě	142

kapitola 6

Kognitivní superschopnosti	146
Funkce a superschopnosti	147
Scénář převratu	152
Moc nad přírodou a moc nad agenty	158

kapitola 7

Superinteligentní vůle	166
Vztah mezi inteligencí a motivací	166
Instrumentální konvergence	171
<i>Sebezáchova</i>	172
<i>Integrita obsahu cílů</i>	172
<i>Vylepšování kognice</i>	175
<i>Zdokonalování technologií</i>	176
<i>Získávání zdrojů</i>	177

kapitola 8

Jak pravděpodobná je zkáza lidstva?	181
Směřovala by intelligenční exploze k existenční katastrofě?	181
Zrádný obrat	183
Druhy zhoubného selhání	188
<i>Zvrácené uskutečnění cíle</i>	188
<i>Překypění infrastruktury</i>	192
<i>Zločiny v mysli</i>	197

kapitola 9

Problém kontroly	199
Dva problémy zastoupení	199
Metody kontroly schopností	202
<i>Metody izolace</i>	202
<i>Metody pobídek</i>	205
<i>Zbrzdění</i>	209
<i>Nástražné mechanismy</i>	213
Metody výběru motivací	215
<i>Přímá specifikace</i>	216
<i>Domestikace</i>	219
<i>Nepřímá normativnost</i>	220
<i>Augmentace</i>	221
Přehled	222

kapitola 10

Věštiny, džinové, suveréní, nástroje	224
Věštiny	224
Džinové a suveréní	229
Nástrojové UI	232
Srovnání	239

kapitola 11

Multipolární scénáře	243
O koních a lidech	244
<i>Mzdy a nezaměstnanost</i>	245
<i>Kapitál a blahobyť</i>	247
<i>Malthusiánský princip v historické perspektivě</i>	249
<i>Populační růst a investice</i>	251
Život v algoritmické ekonomice	253
<i>Dobrovolné otroctví, masová úmrtí</i>	255
<i>Jak zábavná by maximálně efektivní práce byla?</i>	258
<i>Nevědomí subdodavatelé?</i>	261
<i>Evoluce nemusí směřovat vzhůru</i>	264
Utvoří se v porevoluční době globální samovláda?	269
<i>Druhý přechod</i>	269
<i>Superorganismy a úspory z rozsahu</i>	270
<i>Smluvní sjednocení</i>	274

kapitola 12

Získávání hodnot	281
Problém nahrávání hodnot	281
Evoluční výběr	284
Zpětnovazební učení	286
Postupný růst hodnot	287
Motivační lešení	289
Učení se hodnotám	291
Modulace emulace	304
Budování institucí	306
Přehled	312

kapitola 13

Volba kritérií pro volby	315
Potřeba nepřímé normativnosti	315
Koherentní extrapolovaná vůle	318
<i>Několik vysvětlení</i>	319
<i>Proč dát přednost KEV</i>	322
<i>Další poznámky</i>	325
Morální modely	327
Dělej to, co mám na mysli	331
Seznam konstrukčních voleb	334
<i>Obsah cílů</i>	335
<i>Teorie rozhodování</i>	336
<i>Epistemologie</i>	338

<i>Ratifikace</i>	340
Dostatečně dobrý výsledek	342

kapitola 14

Strategická mapa	344
Vědecká a technologická strategie	345
<i>Diferencovaný technologický vývoj</i>	345
<i>Upřednostňované pořadí</i>	347
<i>Rychlost změn a vylepšování kognice</i>	351
<i>Spárování technologií</i>	356
<i>Vypočítavost v komunikaci</i>	359
Cesty a předstupy	361
<i>Dopady hardwarového pokroku</i>	361
<i>Měli bychom podporovat výzkum úplné emulace mozku?</i>	364
<i>Z osobního hlediska dáváme přednost rychlosti</i>	369
Spolupráce	370
<i>Závodová dynamika a její nebezpečí</i>	370
<i>O přínosech kooperace</i>	374
<i>Pracovat společně</i>	380

kapitola 15

Kritické období	383
Filozofie s datem uzávěrky	383
Co bychom měli dělat?	385
<i>Poznání strategického terénu</i>	386
<i>Budování kapacity</i>	387
<i>Konkrétní opatření</i>	388
Ať vyjde na světlo to nejlepší v nás	389

Doslov	391
-------------------------	-----

Poznámky	396
Seznam literatury	473
Seznamy obrázků, tabulek a rámečků	496
Slovníček	498
Poděkování	502
Rejstřík	504

Nedokončená bajka o vrabcích

Bylo právě období stavby hnízd, avšak po dnech dlouhé tvrdé práce teď vrabci seděli ve večerním světle, odpočívali a štěbetali.

„Všichni jsme tak malí a slabí. Pomyslete, jak by byl náš život snadný, kdybychom měli sovu, která by nám pomáhala se stavbou hnízd!“

„Ano!“ řekl další vrabec. „A mohla by se starat o naše staré a o mláďata.“

„Mohla by nám radit a dávat pozor na kočku,“ dodal třetí.

Potom promluvil Pastus, ptačí stařešina: „Vyšleme zvědy do všech stran a pokusme se někde najít opuštěné soví mládě nebo vejce. Třeba by stačila i mladá vrána nebo lasička. Tohle by mohlo být to nejlepší, co nás kdy potkalo, přinejmenším od otevření Pavilonu nevyčerpatelného zrní vedle na dvorku.“

Celé hejno bylo nadšené a všude kolem začali vrabci štěbetat z plných plic.

Jediný Vrzličák, mrzutý jednooký vrabec, nebyl o rozumnosti tohoto podniku přesvědčen. I pravil: „To bude jistě náš konec. Než sovu přivedeme mezi nás, neměli bychom se nejprve zamyslet, jak ji ochočit?“

Pastus odpověděl: „Ochočit sovu, to zní jako neobyčejně náročný úkol. Už nalezení sovího vajíčka bude velmi obtížné. Začněme tedy u toho. Až se nám podaří si sovu opatřit, můžeme začít přemýšlet, jak si poradit s další výzvou.“

„Ten plán má zásadní vadu!“ zapištěl Vrzličák, avšak jeho protesty byly marné, neboť hejno už vzlétlo, aby začalo uskutečňovat Pastovy pokyny.

Pouze dva nebo tři vrabci zůstali na místě. Pokoušeli se společně vymyslet, jak by šlo sovy domestikovat. Brzy si uvědomili, že Pastus

měl pravdu: šlo o neobyčejně náročný úkol, především proto, že jim chyběla skutečná sova, s níž by mohli cvičně pracovat. Přesto se i nadále snažili, jak nejlépe dovedli, a žili ve stálém strachu, že se hejno se sovím vejcem vrátí dříve, než naleznou řešení problému kontroly.

Není známo, jak příběh končí, avšak tuto knihu autor věnuje Vrzličákovi a jeho následovníkům.

Předmluva

Uvnitř vaší lebky se nachází věc, která teď čte tuto větu. Tato věc – lidský mozek – má některé schopnosti, které chybějí mozkům ostatních zvířat. Právě těmto jeho charakteristickým schopnostem vděčíme za své dominantní postavení na planetě Zemi. Jiná zvířata mají silnější svaly a ostřejší drápy, ale my máme chytřejší mozky. Díky své skromné převaze v oblasti obecné inteligence jsme dospěli k vytvoření jazyka, technologie a komplexního společenského uspořádání. Spolu s tím, jak každá generace staví na výtvarcích té předchozí, naše převaha postupem času narůstá.

Pokud jednoho dne vytvoříme strojové mozky s vyšší obecnou inteligencí, než mají mozky lidské, mohla by se tato nová superinteligence stát velice mocnou. A stejně jako osud goril nyní závisí více na nás lidech než na samotných gorilách, závisel by i osud našeho druhu na činech strojové superinteligence.

Máme ovšem jednu výhodu: tato vymoženost by byla naším vlastním dílem. V principu bychom mohli postavit takovou superinteligenci, která by ochraňovala lidské hodnoty. Rozhodně bychom měli silný důvod tak učinit. V praxi ovšem problém kontroly – jak mít pod kontrolou, co superinteligence bude dělat? – vypadá poměrně obtížně. Také se zdá, že dostaneme jenom jednu šanci. Jakmile by nepřátelská superinteligence existovala, zabránila by nám v tom, abychom ji nahradili jinou nebo abychom změnili její preference. Náš osud by byl zpečetěn.

V této knize se pokouším porozumět tomu, jakou výzvu vyhlídka na vznik superinteligence představuje a jak bychom na ni mohli nejlépe odpovědět. Dost možná jde o tu nejdůležitější a nejohroživější

výzvu, jaké kdy lidstvo čelilo. A ať už uspějeme, nebo selžeme, je to pravděpodobně ta poslední výzva, které kdy budeme čelit.

Součástí mé argumentace není tvrzení, že se nacházíme na prahu velkého průlomu ve vývoji umělé inteligence (UI) nebo že můžeme přesně předpovědět, kdy by k němu mohlo dojít. Zdá se celkem pravděpodobné, že nastane někdy v tomto století, ale nevíme to jistě. První dvě kapitoly se zabývají možnými cestami k tomuto průlomu a otázkou, kdy by k němu mohlo dojít. Většina knihy se však týká toho, k čemu dojde poté. Budeme zkoumat kinetiku inteligentní exploze, formy a schopnosti superinteligence a také to, jaké strategické volby budou dostupné superintelligentnímu agentovi, který získá rozhodující převahu. Poté obrátíme pozornost k problému kontroly a budeme se ptát, jak bychom mohli počáteční podmínky zformovat tak, aby výsledný stav umožňoval naše přežití a byl pro nás prospěšný. Ke konci knihy se z většího odstupu podíváme na širší obraz, který před námi na základě našeho zkoumání vyvstane. Předložíme několik návrhů, co bychom v tuto chvíli měli udělat pro zvýšení šancí, že se v budoucnosti vyhneme existenční katastrofě.

Tuto knihu nebylo snadné napsat. Cesta, kterou jsem proklestil, snad ostatním badatelům umožní k nově vymezené hranici dospět rychleji a pohodlněji, aby se tam ocitli svěží a připraveni připojit se k práci na rozšiřování dosahu našeho poznání. (A je-li cesta, kterou jsem připravil, trochu hrbolatá a klikatá, recenzenti snad při posuzování výsledku nepodcení, jak nepřátelským byl terén *ex ante*!)

Tuto knihu nebylo snadné napsat: snažil jsem se, aby se dala snadno číst, ale nemyslím, že se mi to zcela podařilo. Během psaní jsem si coby cílové publikum představoval vlastní mladší já a pokoušel jsem se vytvořit knihu, jakou by mě bavilo číst. Možná se ukáže, že takto určený výsek populace je poněkud úzký. Přesto se domnívám, že by obsah knihy měl být přístupný pro mnoho lidí, pokud nad ním budou přemýšlet a odolají pokušení každou novou myšlenku okamžitě připodobnit tomu klišé ze své kulturní zásobárny, které se jí nejvíce podobá, a tak si ji ihned vyložit nesprávně. Čtenáři postrádající vzdělání v oboru by se neměli nechat odradit matematickými rovnicemi či technickými termíny, jež se v knize

občas objeví, neboť ústřední myšlenky se lze vždy dobrat s pomocí doprovodných vysvětlení. (Čtenáři, kteří se naopak chtějí dozvědět více technických podrobností, jich naleznou dosyta v poznámkách na konci knihy.)¹

Mnohá tvrzení této knihy jsou nejspíš nesprávná.² Je také pravděpodobné, že jsem opomenul vzít v úvahu některé zásadně důležité faktory, a učinil tak neplatnými některé nebo všechny své závěry. Věnoval jsem velkou pozornost tomu, abych v celém textu vyznačil různé stupně nejistoty a jejich nuance, a přeplnil jsem jej nevzhlednými obraty typu „mohlo by“, „může“, „možná“, „zřejmě“, „asi“, „pravděpodobně“. Všechny kvalifikátory byly na svá místa zařazeny pečlivě a uváženě. Avšak tyto místní projevy epistemické skromnosti nejsou postačující; musejí zde být doplněny o systémové přiznání nejistoty a omylnosti. Nejde o falešnou skromnost: jakkoliv se totiž domnívám, že je má kniha pravděpodobně zásadním způsobem nesprávná a zavádějící, myslím si, že alternativní stanoviska představená v existující literatuře jsou podstatně horší – včetně onoho základního stanoviska nebo oné „nulové hypotézy“, podle které je bezpečné či rozumné vyhlídku na vznik superinteligence prozatím ignorovat.

kapitola 1

Dosavadní vývoj a dnešní schopnosti

Nejprve se vrátíme do minulosti. Podíváme-li se na dějiny v nejširším měřítku, uvidíme posloupnost rozdílných režimů růstu, z nichž každý je rychlejší než ten předcházející. Tento vzorec dějinného vývoje podle některých názorů nasvědčuje tomu, že by mohl být možný další, ještě rychlejší režim růstu. Na tento postřeh však neklademe příliš velký důraz – téma této knihy netvoří „zrychlování technologického pokroku“, „exponenciální růst“ ani rozličné myšlenky, jež jsou někdy sdružovány do pojmu „singularita“. Následně přehlédneme dějiny umělé inteligence a poté zmapujeme její dnešní schopnosti. Nakonec se podíváme na některé nedávné průzkumy mínění mezi experty a popřemýšlíme nad naší nevědomostí ohledně doby, kdy dojde k budoucím pokrokům.

Režimy růstu a velké dějiny

Před pouhými pár miliony let se naši předci stále ještě houpali na větvích v afrických pralesích. Vzestup od posledního společného předka lidí a lidoopů k druhu *Homo sapiens* proběhl na geologické i evoluční časové ose velmi rychle. Osvojili jsme si vzpřímené držení těla, získali jsme chápavé palce, a především jsme podstoupili jisté poměrně malé změny ve velikosti mozku a v uspořádání

nervové soustavy, jež vedly k velkému pokroku v našich kognitivních schopnostech. Díky tomu dokážeme mnohem lépe než jakýkoliv jiný druh na planetě abstraktně přemýšlet, předávat si složité myšlenky a hromadit informace napříč generacemi.

Tyto schopnosti lidem umožnily rozvíjet stále účinnější výrobní technologie, takže se naši předci mohli přestěhovat daleko od deštného pralesa a savany. Zvláště po vzniku zemědělství se začala zvyšovat hustota a celková velikost lidské populace. S více lidmi se objevovalo více myšlenek – a díky větší hustotě obyvatelstva se tyto myšlenky mohly šířit rychleji a někteří jednotlivci se mohli věnovat rozvoji specializovaných dovedností. Tento vývoj zvýšil *tempo růstu* ekonomické produktivity a technologické kapacity. K druhému srovnatelnému skokovému navýšení tempa růstu vedly pozdější proměny související s průmyslovou revolucí.

Takové změny v tempu růstu mají významné následky. Před několika stovkami tisíc let, v době rané lidské (či hominidní) prehistorie, probíhal růst tak pomalu, že trvalo zhruba milion let, než se lidská výrobní kapacita zvýšila natolik, aby bylo možné uživit další milion jednotlivců žijících na úrovni existenčního minima. V roce 5000 př. n. l., v době po neolitické revoluci, již bylo tempo růstu o tolik vyšší, že na tentýž objem růstu stačila jenom dvě staletí. V současnosti, v době po průmyslové revoluci, světová ekonomika o tento objem roste v průměru každých devadesát minut.³

Pokud se současné tempo růstu udrží alespoň po středně dlouhou dobu, přinese i ono působivé výsledky. Poroste-li světová ekonomika nadále stejnou rychlostí jako v posledních padesáti letech, bude do roku 2050 svět asi 5,4krát a do roku 2100 přibližně 34krát bohatší, než je dnes.⁴

Avšak perspektiva stále pokračujícího exponenciálního růstu bledne ve srovnání s vyhlídkou na to, co by se stalo, kdyby svět zažil další skokové navýšení *tempa růstu*, jež by svou velikostí bylo srovnatelné s navýšeními spojenými s neolitickou a průmyslovou revolucí. Ekonom Robin Hanson na základě údajů o ekonomice a demografii minulých období odhaduje, že typická doba, za niž se velikost světové ekonomiky zdvojnásobí, činí pro pleistocenní společnost lovců a sběračů 224 000 let, pro společnost zemědělců 909 let a pro

průmyslovou společnost 6,3 roku.⁵ (Současné období podle Hansona modelu představuje směs zemědělského a průmyslového režimu růstu – světová ekonomika jako celek se dosud dvojnásobně za 6,3 roku nezvětšuje.) Pokud by došlo k dalšímu takovému skoku a pokud by byl podobně veliký jako oba předchozí, nastal by tím nový režim růstu, v němž by se světová ekonomika dvojnásobně zvětšovala každé dva týdny.

Z dnešního hlediska působí takové tempo růstu neuvěřitelně. Pozorovatelé z dřívějších epoch by možná považovali za stejně absurdní, že se jednou velikost světové ekonomiky bude zdvojnásobovat několikrát během jednoho lidského života. Právě tuto neobyčejnou situaci přesto nyní považujeme za obyčejnou.

Myšlenka blížící se technologické singularity byla široce zpopularizována, nejprve zásluhou vlivného článku Vernora Vinge a poté zásluhou textů Raye Kurzweila a dalších autorů.⁶ Výraz „singularity“ je však zmateným způsobem používán v mnoha nesourodých významech a byl obdařen hroživou (a přece téměř chiliastickou) aurou technologicko-utopistických konotací.⁷ Protože pro naše účely je většina z těchto významů a konotací irelevantní, můžeme v zájmu větší jasnosti slovo „singularity“ nahradit přesnější terminologií.

Z myšlenek souvisejících se „singularity“ nás zde zajímá možnost *intelenční exploze*, zvláště perspektiva vzniku strojové superintelligence. Takové diagramy růstu, jako jsou ty na obrázku 1, možná některé čtenáře přesvědčí, že je na spadnutí další drastická změna v režimu růstu, srovnatelná s neolitickou nebo průmyslovou revolucí. Tito čtenáři pak mohou dovodit, že lze stěží vymyslet scénář, v němž se doba dvojnásobného růstu světové ekonomiky zkrátí na dva týdny, a jenž přitom nezahrnuje stvoření umělých mozků, které budou mnohem rychlejší a výkonnější než ty dobře známé biologické. Argumentace pro to, že bychom vidinu revoluce ve strojové inteligenci měli brát vážně, se však nemusí opírat o cvičení v prokládání dat křivkou ani o extrapolace z dřívějšího ekonomického růstu. Jak uvidíme, existují silnější důvody, proč bychom se měli mít na pozoru.