NEW YORK TIMES BESTSELLER ★★★ 2014

NICK BOSTROM

SUPERINTELIGENCE

PROSTOR

AŽ BUDOU STROJE CHYTŘEJŠÍ NEŽ LIDÉ

Superinteligence

Vyšlo také v tištěné verzi

Objednat můžete na
www.eprostor.com
www.e-reading.cz
www.palmknihy.cz

Nick Bostrom

Superinteligence – e-kniha


PROSTOR


edice
GLOBUS

Nick Bostrom

Superinteligence

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.

ISBN 978-80-7260-389-3

OBSAH

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 výdobytcích 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 nejhrozivější

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 inteligenční exploze, formy a schopnosti superinteligence a také to, jaké strategické volby budou dostupné superinteligentnímu 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]

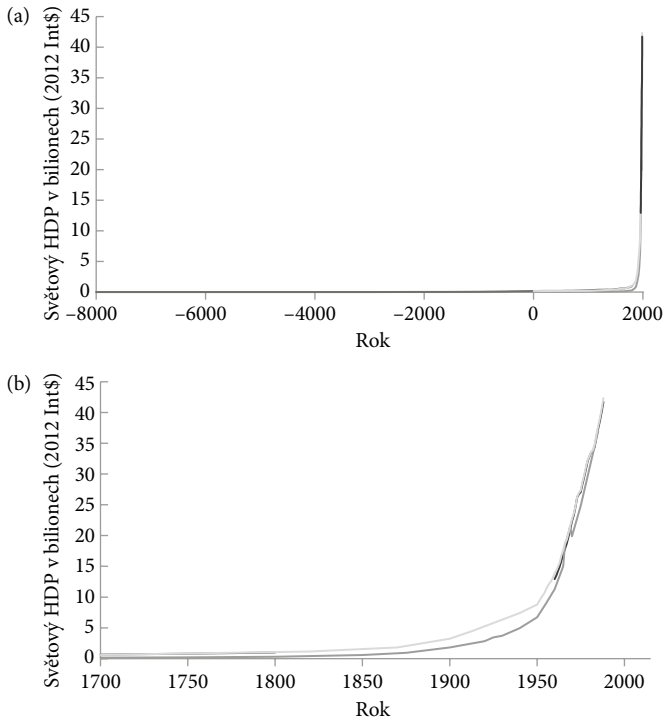
Avšak perspektiva stále pokračujícího exponenciálního růstu bledne ve srovnání s vyhlídkou na to, co by se stalo, kdyby svět zažil další skokové navýšení *tempa růstu*, jež by svou velikostí bylo srovnatelné s navýšeními spojenými s neolitickou a průmyslovou revolucí. Ekonom Robin Hanson na základě údajů 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 Hansonova 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 „singularita" je však zmateným způsobem používán v mnoha nesourodých významech a byl obdařen hrozivou (a přece téměř chiliastickou) aurou technologicko-utopistických konotací.[7] Protože pro naše úvahy je většina z těchto významů a konotací irelevantní, můžeme v zájmu větší jasnosti slovo „singularita" nahradit přesnější terminologií.

Z myšlenek souvisejících se „singularitou" nás zde zajímá možnost *inteligenční exploze*, zvláště perspektiva vzniku strojové superinteligence. 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.



Obrázek 1 Dějiny světového hrubého domácího produktu (HDP) v dlouhodobé perspektivě. Zakreslíme-li je do grafu v lineárním měřítku, vypadají dějiny světové ekonomiky jako vodorovná čára, jež se těsně drží osy *x*, dokud náhle nezamíří prudce nahoru. I pokud si přiblížíme jenom posledních 10 000 let, stále má křivka v podstatě podobu jednoho pravého úhlu (a). Až během posledních asi 100 let se křivka postřehnutelně zvedá nad nulovou úroveň (b). (Různé křivky v grafu odpovídají různým souborům dat, jež dávají lehce odlišné odhady.)[8]

Nadějné vyhlídky

Vznik strojů, jež se nám vyrovnají co do obecné inteligence – tj. jež budou mít zdravý rozum a budou schopny se efektivně učit, uvažovat a plánovat, takže se napříč širokým spektrem přirozených a abstraktních oblastí vyrovnají s komplexními úlohami na zpracovávání informací –, lidé očekávali už od vynálezu počítačů ve čtyřicátých letech

20. století. Tehdy předpovídali, že takové stroje vzniknou zhruba za 20 let.[9] Od té doby se datum jejich očekávaného příchodu vzdalovalo rychlostí jednoho roku za rok, takže i dnešní futurologové, kteří se zabývají možností umělé obecné inteligence, často soudí, že inteligentní stroje se objeví za dvě desetiletí.[10]

Vzdálenost dvou desetiletí je pro hlasatele radikální změny ideální: je dostatečně krátká, aby vzbuzovala pozornost a působila relevantně, ale zároveň dostatečně dlouhá, aby se dalo předpokládat, že během ní dojde k řadě objevů, které si nyní dokážeme představit jenom matně. Srovnejme ji s kratšími časovými obdobími: většina technologií, které budou mít na svět velký dopad v příštích pěti či deseti letech, se už v omezeném rozsahu používá a technologie, které svět přetvoří za patnáct let, pravděpodobně již existují v podobě laboratorních prototypů. Zhruba dvacet let také prognostikovi obvykle zbývá do konce kariéry, což snižuje rizika, která by smělá předpověď mohla představovat pro jeho pověst.

Z toho, že v minulosti byly některé předpovědi ohledně umělé inteligence přehnané, však neplyne, že UI není možná nebo že nikdy nebude vyvinuta.[11] Hlavním důvodem, proč byl pokrok pomalejší, než se očekávalo, byla skutečnost, že technické obtíže spojené s konstrukcí inteligentních strojů byly nakonec větší, než průkopníci UI předvídali. Zůstává však otevřené, jak přesně velké tyto obtíže jsou a jak daleko jsme dnes od toho, abychom je překonali. Občas se ukáže, že problém, jenž na počátku vypadal beznadějně složitě, lze vyřešit překvapivě snadno (ačkoliv k opaku pravděpodobně dochází častěji).

V následující kapitole se podíváme na několik cest, které můžou vést ke strojové inteligenci, jež by se vyrovnala té lidské. Hned na začátku však upozorněme, že ať už se mezi ní a nynějším stavem nachází dalších zastávek, kolik jen chcete, není strojová inteligence lidské úrovně konečnou destinací. Další zastávkou, ležící na dohled od ní, je strojová inteligence, která tu naši převyšuje. Vlak pokroku se možná ve stanici Člověkovice nezastaví, ba v ní ani nezpomalí. Pravděpodobně jí jenom prosviští.

Možná prvním, kdo formuloval základní body tohoto scénáře,

byl matematik I. J. Good, jenž během druhé světové války pracoval jako hlavní statistik v Turingově kryptoanalytickém týmu. V hojně citované pasáži z roku 1965 Good napsal:

> Nechť je ultrainteligentní stroj definován jako takový stroj, který ve všech intelektuálních činnostech dokáže dosáhnout mnohem vyšší úrovně než libovolný člověk, ať jakkoliv chytrý. Protože jednou z lidských intelektuálních aktivit je i navrhování strojů, dokázal by ultrainteligentní stroj navrhovat ještě lepší stroje; bezpochyby by tedy došlo k „inteligenční explozi" a lidská inteligence by zůstala daleko pozadu. První ultrainteligentní stroj je tudíž tím posledním vynálezem, který kdy člověk bude muset vytvořit – pod podmínkou, že tento stroj bude dostatečně povolný, aby nám sdělil, jak jej udržet pod kontrolou.[12]

Dnes se může zdát očividné, že s takovou inteligenční explozí by byla spojena velká existenční rizika, a že bychom se tedy vyhlídkou na ni měli zabývat s tou největší vážností, i pokud by bylo známo (což není), že pravděpodobnost jejího uskutečnění je poměrně malá. Průkopníci v umělé inteligenci však – vzdor svému přesvědčení o brzkém nástupu UI lidské úrovně – většinou nezvažovali možnost nadlidské UI. Jako by své spekulativní svalstvo příliš vyčerpali při koncipování radikální možnosti, že stroje dosáhnou lidské inteligence, takže už nedokázali pochopit, jaký by to nevyhnutelně mělo následek – totiž že by se stroje posléze staly superinteligentními.

Průkopníci oboru UI většinou nepřipouštěli, že by jejich podnik s sebou mohl nést nějaká rizika.[13] Bezpečnostním obavám či etickým pochybnostem týkajícím se stvoření umělých myslí a potenciálních vládců lidstva se nevěnovali ani verbálně, natož aby o nich vážně přemýšleli. Tato mezera je překvapivá i v kontextu dobových standardů kritického hodnocení technologií, jež nebyly právě nejvyšší.[14] Nelze než doufat, že až tento podnik skutečně začne být proveditelný, budeme nejen disponovat technologickou zběhlostí potřebnou k nastartování inteligenční exploze, nýbrž také budeme mít technologie více pod kontrolou. Možná to bude nutné, abychom onen výbuch dokázali přežít.

Než se však obrátíme k tomu, co nás čeká v budoucnosti, bude užitečné se krátce podívat na dosavadní dějiny strojové inteligence.

Období naděje a zoufalství

V létě roku 1956 se deset vědců, které spojoval zájem o neuronové sítě, teorii automatů a zkoumání inteligence, zúčastnilo šestitýdenního pracovního semináře na Dartmouth College. Tento Dartmouthský letní projekt je často považován za událost, která dala vzniknout umělé inteligenci coby výzkumnému oboru. Mnozí z účastníků budou později řazeni mezi jeho zakladatele. Optimismus delegátů se odráží v návrhu předloženém Rockefellerově nadaci, která seminář financovala:

> Navrhujeme, aby bylo provedeno dvouměsíční zkoumání umělé inteligence, jehož se zúčastní deset lidí. … Toto zkoumání bude vedeno domněnkou, že všechny aspekty učení či všechny ostatní rysy inteligence mohou být v principu popsány tak přesně, že bude možné vytvořit stroj, který je bude simulovat. Pokusíme se zjistit, jak vytvořit stroje, které budou užívat jazyk, tvořit abstrakce a pojmy, řešit ty druhy úloh, jež jsou nyní vyhrazeny lidem, a které se budou zdokonalovat. Domníváme se, že když na těchto problémech bude během jednoho léta společně pracovat pečlivě vybraná skupina vědců, bude možné v jednom nebo ve více z nich dosáhnout významného pokroku.

Během šesti desetiletí, která uběhla od tohoto ukvapeného začátku, procházel obor umělé inteligence střídavě obdobími, kdy jej obklopovalo mnoho povyku a velká očekávání, a dobami nezdarů a zklamání.

Podle pozdějšího výroku Johna McCarthyho (hlavního pořadatele dartmouthského semináře) toto první nadšené období, na jehož počátku stálo setkání v Dartmouthu, vystihuje zvolání: „A přece to jde!“ V této rané době badatelé budovali systémy, které měly vyvrátit různá tvrzení typu: „Žádný stroj nikdy nedokáže vykonat *X*.“ Taková skeptická prohlášení byla v té době běžná. Badatelé jim

čelili tak, že vytvářeli malé systémy, které dokázaly *X* provést v „mikrosvětě“ (tj. v dobře definované, omezené oblasti, v níž se mohly v redukované podobě uskutečnit ty výkony, jež měly být předvedeny), čímž prokázali realizovatelnost takových systémů a ukázali, že stroje v principu *X* vykonat můžou. Jeden takový raný systém, nazvaný Logický teoretik (*Logic Theorist*), byl schopen provést důkaz většiny teorémů z druhé kapitoly Whiteheadova a Russellova spisu *Prinicipia Mathematica*, a dokonce objevil jeden důkaz, jenž byl mnohem elegantnější než ten původní. Vyvrátil tak názor, podle kterého stroje mohou „myslet jenom numericky“, a ukázal, že jsou schopny rovněž provádět dedukce a objevovat logické důkazy.[15] Na Logického teoretika navázal program Obecný řešitel úloh (*General Problem Solver*), který v principu dokázal vyřešit celou řadu formálně vymezených úloh.[16] Vznikly rovněž programy schopné řešit takové příklady z infinitezimálního počtu, jaké jsou typické pro první rok vysokoškolského studia, úlohy založené na vizuální analogii (podobné těm, které se objevují v některých IQ testech) a jednoduché algebraické slovní úlohy.[17] Robot *Shakey* (čili Třesavka – tak byl pojmenován pro svůj sklon se během provozu třást) demonstroval, jak lze logické myšlení propojit s vnímáním a použít je k plánování a kontrole fyzické činnosti.[18] Program ELIZA ukázal, jak by počítač mohl napodobit rogeriánského psychoterapeuta.[19] V polovině sedmdesátých let program SHRDLU předvedl, jak se simulovaná robotická ruka dokáže v simulovaném světě geometrických kostek řídit pokyny (a jak program může odpovídat na anglické otázky zadávané uživatelem).[20] V dalších desetiletích byly vytvořeny systémy, které ukázaly, že stroje dokážou komponovat ve stylu různých skladatelů vážné hudby, předčit v některých klinických diagnostických úkolech začínající lékaře, samostatně řídit automobily a přicházet s vynálezy, na které lze získat patent.[21] Objevila se dokonce UI, která vypráví své vlastní vtipy.[22] (Ne že by její humor byl úplně na výši: „Co dostanete, když zkřížíte *citoslovce podivu* a *sýr*? Aj-dam.“ Děti ale svými slovními hříčkami údajně vždy pobaví.)

Často se ukázalo, že metody, které přinesly úspěchy u těchto raných ukázkových systémů, lze obtížně přenést na širší škálu problémů

nebo na náročnější případy daného problému. Jedním z důvodů je „kombinatorická exploze" možností, jež musejí prozkoumat metody založené na něčem takovém jako „prohledávání hrubou silou" (tj. na prověření všech možných variant). Takové metody dobře fungují pro jednoduché případy dané úlohy, avšak selhávají, jakmile se věci začnou trochu komplikovat. Chceme-li například v deduktivním systému, jenž zahrnuje jedno odvozovací pravidlo a pět axiomů, dokázat teorém s důkazem o pěti krocích, můžeme jednoduše provést výčet všech 3125 možných kombinací a u každé z nich ověřit, jestli vede k zamýšlenému závěru. Řešení hrubou silou by fungovalo i pro důkazy o šesti nebo sedmi krocích. Jak se však úkol komplikuje, brzy se tato metoda dostává do nesnází. Dokázat teorém s důkazem o padesáti krocích netrvá jen desetkrát déle než ten, jehož důkaz má pět kroků: uplatníme-li hrubou sílu, ve skutečnosti to vyžaduje, abychom prošli $5^{50} \approx 8{,}9 \times 10^{34}$ možných posloupností – na což nemají dost výpočetního výkonu ani ty nejrychlejší superpočítače.

K překonání kombinatorické exploze jsou zapotřebí algoritmy, jež pracují se strukturou v cílové oblasti a s pomocí heuristického prohledávání, plánování a flexibilních abstraktních reprezentací uplatňují dřívější znalosti, což jsou schopnosti, které byly u prvních systémů UI rozvinuté jen chabě. Výkonnost těchto systémů trpěla také kvůli špatným metodám práce s neurčitostí, závislosti na křehkých a neukotvených symbolických reprezentacích, nedostatku dat a významným hardwarovým omezením (nedostatečné paměťové kapacitě a pomalým procesorům). Povědomí o těchto obtížích vzrůstalo v první polovině sedmdesátých let 20. století. Zjištění, že mnohé projekty v UI nemohou splnit, co původně slibovaly, vedlo k nástupu první „zimy UI": období úspor, kdy klesalo financování, zvyšoval se skepticismus a UI vyšla z módy.

Nové jaro přišlo na začátku osmdesátých let, kdy Japonsko začalo s projektem počítačových systémů páté generace. Šlo o dobře financovaný projekt, na němž spolupracoval veřejný a soukromý sektor a jehož cílem bylo překonat nejlepší UI té doby. Prostředkem k dosažení tohoto cíle mělo být vytvoření architektury masivně paralelních systémů, která by sloužila jako platforma pro umělou

inteligenci. Začátek osmdesátých let byl zároveň dobou, kdy svého vrcholu dosáhla fascinace japonským „poválečným hospodářským zázrakem" a kdy se vůdčí postavy západního obchodu a politiky horlivě snažily odhalit japonský recept v naději, že jeho úspěch zopakují doma. Když se Japonsko rozhodlo investovat do UI, několik dalších zemí je napodobilo.

V následujících letech se velmi rozšířily *expertní systémy*, které měly sloužit jako podpůrné prostředky při rozhodování. Šlo o pravidlové programy, jež činily jednoduché úsudky na základě znalostní báze získané od lidských expertů v dané oblasti, namáhavě ručně zakódované do formálního jazyka. Malé expertní systémy však nebyly dost užitečné a u těch větších se ukázalo, že jejich vývoj, validace a aktualizace jsou velmi nákladné. Navíc se s velkými systémy obvykle těžce pracovalo. Bylo nepraktické opatřovat si samostatný počítač kvůli jedinému programu. Před koncem osmdesátých let i toto plodné období pominulo.

Japonskému projektu počítačů páté generace ani jeho protějškům ve Spojených státech a v Evropě se nepodařilo splnit jejich cíle. Nadešla druhá zima UI. V té době jeden kritik po právu truchlil nad „dosavadními dějinami výzkumu umělé inteligence, které sestávají vždy jen z drobných úspěchů v dílčích oblastech, na něž okamžitě navazuje neúspěch v dosažení obecnějších cílů, na které jako by ony počáteční úspěchy zprvu poukazovaly."[23] Soukromí investoři se začali vyhýbat všem podnikům nesoucím značku „umělá inteligence". Dokonce i mezi akademiky a jejich sponzory se UI stala nevítaným přízviskem.[24]

Technologický vývoj však pokračoval rychlým tempem a v devadesátých letech druhou zimu UI vystřídala obleva. Opětovný optimismus byl vyvolán příchodem nových technik, jež se zdály nabízet alternativy k tradičnímu logicistnímu paradigmatu, často označovanému „Stará dobrá umělá inteligence" či krátce „SDUI" („*Good Old-Fashioned Artificial Intelligence*" – „GOFAI"), které se zaměřovalo na vysokoúrovňovou manipulaci se symboly a svého vrcholu dosáhlo s expertními systémy z osmdesátých let. Nové oblíbené techniky, mezi něž patřily neuronové sítě a genetické algoritmy,

slibovaly překonání některých nedostatků logicistního přístupu, obzvláště „křehkosti", která charakterizovala klasické UI programy (ty obvykle plodily úplné nesmysly, pokud programátoři zadali byť jen jediný lehce nesprávný předpoklad). Nové techniky se mohly pochlubit organičtějším chováním. Například neuronové sítě vykazovaly vlastnost nazývanou „elegantní degradace": výsledkem drobného poškození neuronové sítě obvykle bylo malé zhoršení jejího výkonu, nikoliv úplné selhání. Ještě důležitější bylo, že se neuronové sítě dokázaly učit ze zkušenosti, když objevovaly přirozené způsoby, jak zobecňovat z příkladů, a odkrývaly ve svých vstupních datech skryté statistické vzorce.[25] Díky tomu byly dobré v rozpoznávání vzorů a v klasifikačních úlohách. Pokud například neuronovou síť natrénujeme na souboru dat sestávajícím ze sonarových signálů, můžeme ji naučit, aby akustické profily ponorek, námořních min a mořských živočichů rozeznávala s větší přesností než lidští experti – a to aniž by někdo předem musel přijít na to, jak přesně jednotlivé kategorie definovat nebo jakou váhu přiřadit různým charakteristikám signálů.

Jednoduché neuronové sítě sice byly známy už od konce padesátých let, ale dočkaly se obrody po zavedení algoritmu zpětného šíření, který umožnil trénování vícevrstvých neuronových sítí.[26] Tyto sítě, které mají mezi vstupní a výstupní vrstvou neuronů jednu nebo několik přechodných („skrytých") vrstev, se dokážou naučit mnohem širší škále funkcí než jejich jednodušší předchůdci.[27] V kombinaci se stále výkonnějšími počítači, jež nyní začínaly být dostupné, umožnily tyto vylepšené algoritmy inženýrům tvořit neuronové sítě, které již měly řadu prakticky užitečných aplikací.

Díky vlastnostem, které je přibližovaly lidským mozkům, vycházely neuronové sítě ze srovnání s logickým puntičkářstvím a křehkostí tradičních pravidlových systémů SDUI příznivě – natolik, že zavdaly podnět ke vzniku nového „ismu", totiž *konekcionismu*, který zdůrazňoval důležitost masivně paralelního subsymbolického procesování. Umělé neuronové sítě se od těchto dob staly předmětem více než 150 000 vědeckých studií a i nadále představují důležitou metodu strojového učení.

Dalším přístupem, jehož vznik napomohl k ukončení druhé zimy

UI, jsou evoluční metody, například genetické algoritmy a genetické programování. Ty možná měly v akademické sféře menší vliv než neuronové sítě, avšak byly široce zpopularizovány. U evolučních modelů udržujeme populaci možných řešení (kterými mohou být datové struktury nebo programy) a nahodile generujeme nová potenciální řešení tak, že členy původní populace vystavíme mutacím nebo je rekombinujeme. V pravidelných intervalech populaci prořezáváme aplikací selekčního kritéria (kriteriální či „fitness" funkce), které dovoluje přejít do dalšího pokolení jenom těm lepším řešením. Pokud tento postup zopakujeme přes tisíce generací, průměrná kvalita řešení přítomných v populaci se bude postupně zvyšovat. Když tento druh algoritmu funguje, dokáže dát vzniknout účinným řešením nejrůznějších úloh – řešením, která mohou být pozoruhodně novátorská a neintuitivní a která se často podobají spíše přírodním útvarům než něčemu, co by vyprojektovali lidští inženýři. K tomu přitom v principu může dojít i bez toho, aby člověk programu dodával mnoho dalších vstupů kromě počátečního vymezení kriteriální funkce, která je často velice jednoduchá. V praxi je však ke správnému fungování evolučních metod zapotřebí velkých dovedností a vynalézavosti, zvláště při koncipování dobrého formátu reprezentace. Bez účinné metody kódování možných řešení (bez genetického jazyka odpovídajícího latentní struktuře v cílové oblasti) většinou evoluční hledání donekonečna bloudí v rozlehlém prohledávaném prostoru nebo se zasekne na lokálním optimu. A dokonce i v případě, že dobrý formát reprezentace nalezneme, je evoluční algoritmus výpočetně náročný a často podléhá kombinatorické explozi.

Neuronové sítě a genetické algoritmy jsou příklady metod, které v devadesátých letech probudily nadšení, protože se zdály nabízet alternativy ke stagnujícímu paradigmatu SDUI. Naším záměrem zde však není vychvalovat tyto dvě metody nebo je vyvyšovat nad množství ostatních technik strojového učení. Jedním z klíčových teoretických posunů posledních dvaceti let ve skutečnosti bylo jasnější poznání, že různé techniky, jež na povrchu vypadají různorodě, lze pochopit jako zvláštní případy uvnitř společného matematického rámce. Mnohé typy neuronových sítí například můžeme vnímat jako

klasifikátory, které vykonávají zvláštní druh statistického výpočtu (metodu maximální věrohodnosti).[28] To pak umožňuje neuronové sítě srovnávat s rozsáhlejší skupinou algoritmů pro učení klasifikátorů na základě příkladů. Do té patří mimo jiné „rozhodovací stromy", „logistické regresní modely", „metody podpůrných vektorů", „prosté bayesovské klasifikátory" či „algoritmus *k*-nejbližších sousedů".[29] Podobně lze genetické algoritmy chápat jako zvláštní případy stochastických gradientních algoritmů, jež jsou zase podmnožinou širší množiny optimalizačních algoritmů. Každý z těchto algoritmů pro budování klasifikátorů či pro prohledávání prostoru řešení má své výhody a nevýhody, které mohou být matematicky zkoumány. Algoritmy se od sebe liší svými požadavky na procesorový čas a paměťový prostor, tím, s jakými induktivními předpoklady pracují, jak snadno do nich může být zahrnut externě vytvořený obsah a jak průhledné je jejich vnitřní fungování pro lidského analytika.

Všechen ten lesk strojového učení a tvořivého řešení problémů tedy ukrývá skupinu přesně matematicky vymezených kompromisů. Ideálem je optimální bayesovský agent, který využívá dostupných informací způsobem, jenž je z hlediska teorie pravděpodobnosti nejlepší možný.* Tento ideál je nedosažitelný, neboť je příliš výpočetně náročný na to, aby mohl být implementován v jakémkoliv skutečném počítači (viz rámeček 1). V souladu s tím můžeme výzkum v umělé inteligenci chápat jako pátrání po zkratkách: po schůdných cestách, jimiž bychom se přiblížili bayesovskému ideálu tak, že obětujeme něco z optimálnosti nebo obecnosti při zachování vysokého výkonu ve skutečných oblastech našeho zájmu.

Odrazem právě vykresleného obrazu je práce, jež byla v posledních dekádách vykonána v oblasti pravděpodobnostních grafických

* Jako „agent" se v oboru umělé inteligence označuje každá entita, která dokáže vnímat své prostředí (prostřednictvím svých „senzorů") a působit na ně (prostřednictvím svých „akčních členů"); agentem tedy může být jak jednoduchý robot, tak systém UI nebo například člověk. V této knize se „agenty" většinou budou rozumět „umělí inteligentní agenti", tedy uměle stvořené entity „se senzory a akčními členy, které si dokážou vytvářet reprezentace znalostí, činit úsudky a jednat" (viz autorovu poznámku 163 v kapitole 2). – Pozn. překl.

Rámeček 1 **Optimální bayesovský agent**

Ideální bayesovský agent vychází z prvotního „apriorního rozdělení pravděpodobnosti" („prioru") – funkce, která každému z možných světů (tj. každému maximálně specifickému obrazu o tom, jak se to se světem může mít) přisuzuje určitou pravděpodobnost.[30] Toto rozdělení zahrnuje nějaký „induktivní předpoklad" (*inductive bias*) toho druhu, jako že jsou vyšší stupně pravděpodobnosti připisovány jednodušším možným světům. (Jedním ze způsobů, jak jednoduchost možného světa formálně definovat, je stanovení jeho „Kolmogorovovy složitosti", která odpovídá délce nejkratšího počítačového programu, jenž generuje úplný popis daného světa.)[31] Zahrnuje rovněž jakékoliv výchozí znalosti, které chce programátor agentovi poskytnout.

Když agent ze svých senzorů získá nové informace, aktualizuje rozdělení pravděpodobnosti tak, že je těmito novými informacemi podmíní podle Bayesovy věty.[32] Podmiňování je matematická operace, která uděluje nulovou pravděpodobnost těm možným světům, které jsou neslučitelné se získanými informacemi, a renormalizuje pravděpodobnostní rozdělení přes zbývající možné světy. Výsledkem je „aposteriorní pravděpodobnostní rozdělení" („posterior"), které agent může v dalším kroku použít jako nový prior. Spolu s tím, jak agent činí další pozorování, koncentruje se gros pravděpodobnosti do zmenšující se množiny možných světů, které jsou i nadále v souladu s pozorovanými daty; a z těchto možných světů vždy mají vyšší pravděpodobnost ty jednodušší.

Pro ilustraci si pravděpodobnost můžeme představit jako písek nasypaný na velkém listu papíru. Papír je rozdělen na několik různě velkých částí, z nichž každá odpovídá jednomu možnému světu, přičemž ty větší z nich odpovídají jednodušším světům. Představme si nyní vrstvu písku rovnoměrně rozloženou po celém papíru: to je naše apriorní pravděpodobnostní rozdělení. Kdykoliv je učiněno pozorování, se kterým některé možné světy nejsou slučitelné, odstraníme písek z odpovídajících částí papíru a spravedlivě jej rozdělíme mezi zbývající oblasti. Celkové množství písku na papíru se tedy nemění, jenom se spolu s tím, jak se hromadí data z pozorování, koncentruje

v menším množství oblastí. Tento obraz znázorňuje učení v jeho nejčistší podobě. (Chceme-li zjistit pravděpodobnost nějaké *hypotézy*, jednoduše spočítáme množství písku přítomného ve všech částech papíru, které odpovídají těm možným světům, v nichž je daná hypotéza pravdivá.)

Prozatím jsme definovali jenom pravidlo učení. Abychom získali agenta, musíme definovat rovněž pravidlo rozhodování. Za tím účelem agentovi udělíme „užitkovou funkci", která každému možnému světu přiřadí nějaké číslo, jež bude vyjadřovat, jak žádoucí tento svět je podle základních preferencí agenta. Nyní bude agent v každém kroku volit jednání s nejvyšším očekávaným užitkem.[33] (Aby toto jednání objevil, může agent sestavit seznam všech možných jednání. Poté může vypočítat pravděpodobnostní rozdělení podmíněné určitým jednáním – to rozdělení pravděpodobnosti, k němuž bychom dospěli, pokud by současné pravděpodobnostní rozdělení bylo podmíněno pozorováním, že se toto jednání uskutečnilo. A nakonec může spočítat očekávanou hodnotu tohoto jednání: tou je součet hodnot všech možných světů vynásobených pravděpodobnostmi, které tyto světy mají v pravděpodobnostním rozdělení podmíněném daným jednáním.)[34]

Pravidlo učení a pravidlo rozhodování společně pro daného agenta definují jeho „pojem optima". (V zásadě identický pojem optima se používá v celé řadě oblastí od umělé inteligence přes epistemologii a filozofii vědy až po ekonomii či statistiku.)[35] Ve skutečnosti není možné takového agenta vytvořit, protože provedení potřebných výpočtů přesahuje schopnosti našich počítačů. Každý pokus o jejich vykonání podléhá právě takové kombinatorické explozi, jakou jsme popisovali v souvislosti se SDUI. Abychom pochopili, proč tomu tak je, uvažme jednu maličkou podmnožinu všech možných světů: představme si možné světy sestávající z jediného počítačového monitoru plovoucího v nekonečném vakuu. Tento monitor má 1000×1000 pixelů, z nichž každý je buď neustále zapnutý, nebo neustále vypnutý. Dokonce i tato podmnožina možných světů je ohromně veliká: počet možných stavů monitoru – těch je $2^{(1000 \times 1000)}$ – převyšuje předpokládané množství

výpočtů, k nimž kdy dojde v pozorovatelném vesmíru. Nedokážeme tedy ani udělat výčet všech možných světů z této maličké podmnožiny možných světů, natož abychom mohli u každého z nich provozovat nějaké složitější výpočty.

Ač nejsou fyzicky uskutečnitelné, mohou mít pojmy optima teoretický význam. Poskytují nám standard, kterým můžeme poměřovat jejich heuristické aproximace. Někdy můžeme také přemýšlet, jak by se optimální agent zachoval v nějakém konkrétním případě. S některými dalšími pojmy optima pro umělé agenty se setkáme v kapitole 12.

modelů, jakými jsou například bayesovské sítě. Bayesovské sítě umožňují výstižně zobrazovat vztahy stochastické a podmíněné nezávislosti, které platí v nějaké dílčí oblasti. (Využití takovýchto vztahů má zásadní význam pro překonání kombinatorické exploze, která je pro pravděpodobnostní usuzování stejně velkým problémem jako pro logickou dedukci.) Zároveň nám poskytují důležité vhledy do pojmu kauzality.[36]

Když problémy týkající se učení, s nimiž se setkáváme v různých dílčích oblastech, uvedeme do vztahu s obecným problémem bayesovského usuzování, má to mimo jiné tu výhodu, že objev nových algoritmů, které činí bayesovské usuzování efektivnějším, ihned vede ke zlepšením napříč mnoha různými oblastmi. Například pokroky v aproximačních technikách „Monte Carlo" byly přímo uplatněny v počítačovém vidění, robotice a výpočetní genomice. Další výhodou je, že tak badatelé v různých oborech mohou snadněji sdílet své objevy. Grafické modely a bayesovská statistika se staly společným předmětem výzkumu v mnoha oborech, například ve strojovém učení, statistické fyzice, bioinformatice, kombinatorické optimalizaci a v teorii komunikace.[37] K leckterým nedávným pokrokům ve strojovém učení došlo díky uplatnění formálních objevů, k nimž se původně dospělo v jiných oborech. (Praktické aplikace strojového učení těží také z rychlejších počítačů a větší dostupnosti rozsáhlých souborů dat.)

Současný stav oboru

Už v současné době umělá inteligence v lecčems lidskou inteligenci překonává. Tabulka 1 mapuje herní dovednosti dnešních počítačů a ukazuje, že v celé řadě her nyní umělá inteligence poráží nejlepší lidské hráče.[38]

Tabulka 1 ***Herní dovednosti UI***

Dáma	Nadlidská úroveň	Prvním programem, který se nějakou hru naučil hrát lépe než jeho stvořitel, byl program na hraní dámy od Arthura Samuela, původně napsaný roku 1952 a později vylepšený (do verze z roku 1956 byly zapracovány techniky strojového učení).[39] V roce 1994 program CHINOOK porazil úřadujícího lidského mistra světa; to bylo poprvé, kdy počítačový program vyhrál oficiální světový šampionát v dovednostní hře. V roce 2002 Jonathan Schaeffer se svým týmem dámu „vyřešil", tj. vytvořil program, který vždy táhne tím nejlepším možným způsobem (program pracuje s metodou alfa-beta prořezávání v kombinaci s databází 39 bilionů koncovek hry). Hrají-li dokonale oba hráči, končí hra remízou.[40]
Vrhcáby	Nadlidská úroveň	1979: Program BKG, vytvořený Hansem Berlinerem, porazil v exhibičním zápase mistra světa ve vrhcábech. Šlo o první počítačový program, který v nějaké hře porazil světového šampiona. Berliner však později vítězství přičítal šťastným hodům kostkami.[41] 1992: Program TD-Gammon od Gerryho Tesaura dosáhl úrovně mistrů světa s využitím metody temporální diference (druhu zpětnovazebního učení) a opakovaných her proti sobě samému za účelem sebezlepšování.[42] Od té doby již vrhcábové programy nejlepší lidské hráče dalece předčily.[43]
Traveller TCS	Nadlidská úroveň (ve spolupráci s člověkem)[44]	V letech 1981 a 1982 program Eurisko od Douglase Lenata vyhrál mistrovství Spojených států ve futuristické hře simulující námořní válku Traveller TCS. V důsledku toho byly provedeny změny v herních pravidlech, jež měly znemožnit nekonvenční strategie využívané programem.[45] Eurisko disponoval heuristikou k budování své flotily a také heuristikou k úpravě své heuristiky.

Othello	Nadlidská úroveň	1997: Program Logistello vyhrál všechny hry v šestikolovém utkání se světovým šampionem Takešim Murakamim.[46]
Šachy	Nadlidská úroveň	1997: Počítač Deep Blue porazil Garriho Kasparova, mistra světa v šachu. Kasparov tvrdí, že v některých tazích počítače zahlédl záblesky skutečné inteligence a tvořivosti.[47] Od té doby se šachové programy nadále zlepšují.[48]
Křížovky	Expertní úroveň	1999: Program Proverb překonal průměrného křížovkáře.[49] 2012: Program Dr. Fill, vytvořený Mattem Ginsbergem, se umístil mezi nejlepší čtvrtinou (jinak lidských) účastníků Amerického křížovkářského turnaje. (Výkony Dr. Filla byly nevyrovnané. Dokázal vyluštit křížovky, které lidé hodnotili jako nejobtížnější, ale vyvedly jej z míry dvě nestandardní úlohy, v nichž bylo zapotřebí psát odpovědi pozpátku nebo diagonálně.)[50]
Scrabble	Nadlidská úroveň	Od roku 2002 software na hraní scrabblu překonává nejlepší lidské hráče.[51]
Bridž	Úroveň nejlepších lidských hráčů	V roce 2005 už se software na hraní bridže vyrovnal nejlepším lidským hráčům.[52]
Jeopardy!	Nadlidská úroveň	2010: Počítač Watson od IBM porazil dva nejlepší hráče všech dob Kena Jenningse a Brada Ruttera.[53] Jeopardy! je televizní soutěž, v níž účastníci odpovídají na otázky z historie, literatury, sportu, zeměpisu, popkultury, vědy a jiných oblastí. Otázky mají formu vodítek a často pracují se slovními hříčkami. (Přibližnou obdobou této hry bylo české Riskuj! – Pozn. překl.)
Poker	Různorodá úroveň	Ve full-ring hrách Texas hold'em pokeru počítačové programy stále zaostávají za nejlepšími lidskými hráči, avšak v některých jiných variantách dosahují nadlidské úrovně.[54]
FreeCell	Nadlidská úroveň	Program, který používá heuristiku vyvinutou s pomocí genetických algoritmů, v této variantě pasiánsu (jež je ve své zobecněné formě NP-úplná) poráží špičkové lidské hráče.[55]

Go	Úroveň velmi dobrého amatéra	V roce 2012 byly programy série Zen v rychlé variantě hry na úrovni třídy 6. dan (tj. na úrovni velmi dobrých amatérských hráčů). Tyto programy využívají prohledávání stromu metodou Monte Carlo a technik strojového učení.[56] Programy na hraní Go se zlepšují zhruba tempem jednoho danu za rok; pokud se toto tempo udrží, mohly by lidského šampiona porazit asi za deset let.

Tyto úspěchy dnes možná nevypadají nijak zvlášť působivě, avšak jen proto, že naše standardy se úměrně dosahovaným pokrokům neustále zvyšují. Kupříkladu mistrovství ve hře v šachy bylo svého času chápáno jako ztělesnění lidských rozumových schopností. Podle slov několika expertů z padesátých let: „Kdybychom dokázali vynalézt úspěšný stroj na hraní šachů, pronikli bychom, jak se zdá, k samotnému jádru lidské intelektuální činnosti.“[57] Dnes už se nezdá, že by tomu tak bylo. Lze mít pochopení pro Johna McCarthyho, který hořekoval, že „jakmile něco začne fungovat, lidé to přestanou nazývat umělou inteligencí“.[58]

V jednom důležitém smyslu se však šachový program ukázal být menším triumfem, než mnozí očekávali. Kdysi se předpokládalo – možná ne bez dobrých důvodů –, že počítač by se v šachu mohl vyrovnat velmistrům jen tehdy, pokud by byl obdařen vysokou *obecnou* inteligencí.[59] Lidé se například možná domnívali, že šachové mistrovství vyžaduje schopnost naučit se abstraktním pojmům, chytře promýšlet strategii, pružně plánovat, činit celou řadu důvtipných logických dedukcí, a snad dokonce modelovat myšlení protivníka. Není tomu tak. Jak vyšlo najevo, zcela vyhovující šachový program lze vybudovat na základě jednoúčelového algoritmu.[60] Na počítačích s rychlými procesory, které začaly být dostupné ke konci 20. století, dokážou tyto programy hrát velmi dobře. Takováto UI má však úzké zaměření: hraje šachy, nic víc neumí.[61]

V jiných oblastech se naopak řešení ukázala být složitější a pokrok pomalejší, než se původně předvídalo. Informatik Donald Knuth jednou vyjádřil svůj podiv, že „UI nyní dokáže v zásadě všechno, k čemu je zapotřebí ‚myšlení‘, ale nedokáže dělat většinu věcí, které

lidé a zvířata dělají ‚bezmyšlenkovitě' – to je z nějakého důvodu mnohem těžší!"[62] Analýza vizuálních dat, rozpoznávání předmětů či kontrola robotova chování během jeho interakce s přirozeným prostředím se ukázaly jako náročné problémy. I zde nicméně došlo a nadále dochází k solidním pokrokům, kterým napomáhá také stálé zdokonalování hardwaru.

Jako obtížně dosažitelné cíle se projevily také zdravý rozum a porozumění přirozenému jazyku. Často se dnes soudí, že dosažení lidské úrovně v těchto úlohách je „UI-úplným" (*AI-complete*) problémem, tedy že vyřešit tyto problémy je v podstatě stejně obtížné jako vybudovat stroje, jejichž inteligence se po všech stránkách vyrovná té lidské.[63] Jinak řečeno, pokud by se někomu vskutku podařilo vytvořit UI, která by přirozenému jazyku rozuměla stejně dobře jako dospělý člověk, tou dobou by se vší pravděpodobností buď již měl vytvořenou UI, která se lidské inteligenci vyrovná i ve všem ostatním, anebo by byl tomuto cíli velmi blízko.[64]

Jak se ukázalo, mistrovství ve hraní šachu lze dosáhnout s pomocí překvapivě jednoduchého algoritmu. Je lákavé spekulovat, že by prostřednictvím nějakého nečekaně jednoduchého algoritmu šlo získat i jiné schopnosti, například obecnou schopnost usuzovat nebo nějakou klíčovou dovednost využívanou při programování. Z toho, že v danou chvíli nejlepší výkony předvádí nějaký složitý mechanismus, ještě neplyne, že stejné nebo lepší výsledky nemůže mít žádný jednoduchý mechanismus. Možná jsme tu jednodušší alternativu prostě zatím neobjevili. Ptolemaiovský systém (podle kterého Země stojí ve středu vesmíru a Slunce, Měsíc, planety a hvězdy obíhají kolem ní) byl po více než tisíc let tím nejmodernějším astronomickým modelem a jeho prediktivní přesnost byla během staletí zlepšována tím, že byl postupně komplikován: k postulovaným nebeským pohybům byly přidávány další a další epicykly. Potom celý tento systém nahradila Koperníkova heliocentrická teorie, která byla jednodušší a měla (byť až poté, co ji dále rozpracoval Kepler) větší prediktivní přesnost.[65]

Umělá inteligence se nyní používá v tolika oblastech, že by nemělo smysl zde podávat jejich úplný přehled. Zmiňme však alespoň

několik příkladů, abychom získali představu o šíři jejích aplikací. Vedle UI určených ke hraní her, vyjmenovaných v tabulce 1, existují sluchadla vybavená algoritmy, jež dokážou odfiltrovat okolní šum; navigační zařízení, která zobrazují mapy a radí řidičům; doporučovací systémy, které uživateli doporučují knihy a hudební alba na základě toho, jaké zboží si dříve koupil a jak je ohodnotil; klinické systémy pro podporu rohodování, jež lékařům pomáhají s diagnózou rakoviny prsu či s interpretací elektrokardiogramů a navrhují léčebné plány. Existují robotičtí domácí mazlíčci a uklízecí roboti, roboti na sekání trávy, záchranní roboti, chirurgičtí roboti a více než milion průmyslových robotů.[66] Světová populace robotů přesahuje 10 milionů.[67]

Moderní programy na rozpoznávání řeči založené na statistických technikách (jako jsou skryté Markovovy modely) jsou nyní dostatečně přesné pro praktické využití; hrubá verze některých částí této knihy byla napsána s jejich pomocí. Osobní digitální asistenti, například Siri od Apple, reagují na vyřčené pokyny, zodpovídají jednoduché dotazy a provádějí příkazy. Optické rozpoznávání znaků v ručně i strojově psaných textech se běžně používá při třídění pošty a při digitalizaci starých dokumentů.[68]

Strojové překlady jsou stále nedokonalé, ale pro mnohé praktické aplikace postačují. Staré překladače, jež se držely paradigmatu SDUI, pracovaly s ručně naprogramovanými mluvnicemi, které museli kvalifikovaní lingvisté vybudovat od nuly, a to pro každý jazyk zvlášť. Novější systémy využívají techniky statistického strojového učení, které na základě pozorovaného jazykového úzu automaticky vytvářejí statistické modely. Parametry pro tyto modely stroje vyvozují z analýzy bilingvních korpusů. Tento přístup se obejde bez lingvistů: programátoři vytvářející tyto systémy ani nemusejí umět jazyky, s nimiž pracují.[69]

Technologie na rozpoznávání obličejů se v posledních letech zlepšily natolik, že jsou nyní používány na automatizovaných hraničních přechodech v Evropě a Austrálii. Ministerstvo zahraničí Spojených států pro zpracovávání žádostí o víza provozuje systém na rozpoznávání tváří, v jehož databázi je více než 75 milionů fotografií.

Sledovací systémy využívají stále sofistikovanější umělou inteligenci a technologie dolování dat k analýze hlasových záznamů, videí či textů, které jsou ve velkém sbírány z médií elektronické komunikace a skladovány v obřích datových centrech.

Programy na dokazování vět a řešení rovnic jsou nyní tak dobře etablované, že už se v podstatě ani nepovažují za UI. Řešitele rovnic tvoří součást vědeckých programů na matematické výpočty, jako je *Mathematica*. Výrobci čipů běžně využívají metod formální verifikace včetně strojového dokazování vět, aby si před zahájením výroby ověřili chování navržených obvodů.

Americká armáda a tajné služby jako první začaly v širokém měřítku nasazovat roboty na zneškodňování bomb, pozorovací a útočné drony a další bezpilotní stroje. Ty jsou z většiny stále odkázány na dálkové řízení, avšak pracuje se na rozšíření jejich autonomních schopností.

Významných úspěchů bylo dosaženo v oblasti inteligentního rozvrhování. Během operace Pouštní bouře v roce 1991 byl program DART, nástroj pro automatizované plánování a rozvrhování logistiky, využit s takovými výsledky, že podle tvrzení DARPA (Agentury ministerstva obrany pro pokročilé výzkumné projekty) se její třicetiletá investice do UI vrátila i s úroky už díky této jediné aplikaci.[70] Aerolinie ve svých rezervačních systémech uplatňují důmyslné systémy na rozvrhování a na určování cen. Mnohé firmy využívají umělou inteligenci v systémech pro kontrolu stavu zásob. Používají také automatické telefonické rezervační systémy a zákaznické linky připojené k softwaru na rozpoznávání hlasu, které jejich nešťastné zákazníky provádějí nepřehledným bludištěm nabízených voleb.

Technologie UI stojí v základu mnohých internetových služeb. Software dohlíží na světový e-mailový provoz a bayesovským spamovým filtrům se – navzdory stálému úsilí spammerů obelstít opatření, která jsou proti nim uplatňována – z větší části podařilo udržet záplavu spamu v patřičných mezích. Software, jenž využívá UI komponenty, je zodpovědný za automatické schvalování nebo odmítání transakcí provedených kreditní kartou a neustále sleduje aktivitu na bankovních účtech, aby odhalil známky podvodů. Také systémy

pro získávání informací využívají strojové učení ve velkém rozsahu. Webový vyhledávač Google je možná tím nejlepším systémem UI, jaký byl dosud vytvořen.

Musíme zdůraznit, že neexistuje ostrá dělicí čára mezi umělou inteligencí a softwarem obecně. Některé z výše uvedených aplikací lze považovat spíše za aplikace generického softwaru než konkrétně za aplikace umělé inteligence – to nás však přivádí zpět k McCarthyho výroku, že jakmile něco začne fungovat, přestane se to nazývat umělou inteligencí. Pro naše účely má větší význam rozlišení mezi systémy s úzkým rozsahem kognitivních schopností (ať už jsou nazývány „UI", nebo ne) a systémy, jejichž schopnost řešit úlohy má širší uplatnění. V podstatě všechny systémy, které se v současnosti používají, jsou prvního druhu. Mnohé z nich však zahrnují složky, jež by se také mohly stát součástí budoucí obecné umělé inteligence nebo by mohly být užitečné při jejím rozvoji: těmi jsou například klasifikátory, vyhledávací algoritmy, plánovače, řešitele a reprezentační rámce.

Jedním důležitým a nesmírně konkurenčním prostředím, v němž dnes UI působí, jsou globální finanční trhy. Významné investiční společnosti ve velkém využívají automatizované systémy pro obchodování na burze. Některé z těchto systémů prostě jenom automatizují provedení prodejních nebo nákupních příkazů vydaných lidským správcem fondu, jiné však sledují složité obchodní strategie, které se přizpůsobují měnícím se tržním podmínkám. Analytické systémy s pomocí rozmanitých technik dolování dat a analýz časových řad zkoumají trhy s cennými papíry, aby odhalily vzorce a trendy jejich vývoje a aby nalezly korelace mezi historickými pohyby cen a vnějšími proměnnými, například klíčovými slovy ve zprávách. Poskytovatelé finančního zpravodajství prodávají informační kanály speciálně přizpůsobené potřebám této UI. Jiné systémy se specializují na hledání vhodných příležitostí pro arbitráže (ať v rámci jednoho, nebo mezi dvěma různými trhy) či na vysokofrekvenční obchodování, které se snaží těžit z nepatrných pohybů cen, k nimž dochází v řádu milisekund (v tomto časovém měřítku získává význam dokonce i zpoždění signálů přenášených rychlostí světla v optických

vláknech, takže je výhodné počítače umístit poblíž burzy). Za více než polovinu obchodů s kmenovými akciemi na amerických trzích zodpovídají algoritmičtí vysokofrekvenční obchodníci.[71] Algoritmické obchodování bylo zapleteno do burzovního krachu „Flash Crash", k němuž došlo v roce 2010 (viz rámeček 2).

Rámeček 2 „Flash Crash" roku 2010

Šestého května 2010 stačily americké akciové trhy do odpoledních hodin klesnout o 4 % kvůli obavám vyvolaným evropskou dluhovou krizí. Ve 14.32 velký prodejce (rodina podílových fondů) spustil prodejní algoritmus, který měl prodávat vysoké množství e-mini S&P 500 futures kontraktů, a to rychlostí navázanou na míru likvidity na burze (monitorovanou minutu po minutě). Tyto kontrakty nakupovali algoritmičtí vysokofrekvenční obchodníci, kteří byli naprogramováni tak, aby kontrakty rychle prodali dalším obchodníkům a vystoupili tak ze své dočasné dlouhé pozice. Protože poptávka hlavních kupujících byla nízká, začali algoritmičtí obchodníci kontrakty prodávat hlavně jiným algoritmickým obchodníkům, kteří je předávali zase dalším algoritmickým obchodníkům. Vznikl tak efekt „horké brambory", jehož následkem se zvýšil objem obchodů – a to prodejní algoritmus interpretoval jako ukazatel vysoké likvidity, což jej podnítilo ke zvýšení tempa, jímž uváděl e-mini kontrakty na trh. Tím se dále roztáčela sestupná spirála. Od jistého bodu se vysokofrekvenční obchodníci začali stahovat z trhu a snižovat tak likviditu; mezitím ceny dále klesaly. Ve 14.45 bylo obchodování s e-mini kontrakty zastaveno automatickým burzovním „jističem". Když bylo o pouhých pět sekund později obnoveno, ceny se stabilizovaly a brzy se začaly vracet na původní úroveň. Ale na vrcholu krize z trhu na chvíli zmizel bilion dolarů – a vedlejší důsledek byl, že značné množství obchodních transakcí s jednotlivými cennými papíry bylo provedeno za „nesmyslné" ceny, třeba za jeden cent nebo 100 000 dolarů. Poté co bylo obchodování pro ten den uzavřeno, sešli se zástupci burz s regulačními orgány a rozhodli se zrušit všechny transakce provedené za ceny, jež se o 60 nebo více procent vzdálily od předkrizových hodnot (usoudili,

že takové transakce byly „zjevně chybné" a lze je tedy podle existujících obchodních pravidel *post facto* zrušit).[72]

Převyprávění této epizody je pro nás odbočkou, protože počítačové programy, které do ní byly zapletené, nebyly nijak zvlášť inteligentní ani důmyslné a protože typ hrozby, který přinesly, se zásadně liší od rizik spojených se strojovou superinteligencí, na něž upozorníme v dalších částech knihy. Přesto tyto události mohou ilustrovat několik užitečných ponaučení. Zaprvé slouží jako připomínka, že souhra mezi složkami, z nichž každá je sama o sobě jednoduchá (například mezi prodejním algoritmem a programy na vysokofrekvenční algoritmické obchodování), může přinést složité a neočekávané následky. Jak jsou do systému zaváděny nové prvky, může v něm docházet k nárůstu systémového rizika, jež nemusí být patrné, dokud se něco nepokazí (a někdy ani potom ne).[73]

Dalším ponaučením je, že když chytří profesionálové udělí programu instrukce založené na předpokladu, který působí rozumně a za normálních okolností je správný (například že objem obchodování je dobrým ukazatelem likvidity trhu), může to přivodit katastrofické následky, pokud se program touto instrukcí s železnou logickou konzistencí nadále řídí i v neočekávané situaci, v níž se daný předpoklad ukáže jako neplatný. Algoritmus prostě dělá to, co dělá – a nejde-li o velmi zvláštní druh algoritmu, nezáleží mu na tom, že se chytáme za hlavu a lapáme po dechu v němé hrůze nad nevhodností jeho konání. S tímto tématem se ještě setkáme.

Třetí poznámka k burzovnímu krachu je ta, že automatizace k němu sice přispěla, napomohla však zároveň k jeho vyřešení. Předprogramovaný jistič, který pozastavil obchodování, jakmile ceny začaly být příliš divoké, měl nařízeno tak učinit automaticky, protože se správně předjímalo, že k rozhodujícím událostem může dojít tak rychle, že lidé nestihnou včas zareagovat. Tento požadavek, abychom se spíše než na lidský dozor během chodu programu spoléhali na předinstalovanou a automatickou bezpečnostní funkci, opět předznamenává jeden motiv, který bude v našem zkoumání strojové superinteligence hrát významnou roli.[74]

Názory na budoucnost strojové inteligence

Díky pokrokům na dvou frontách – vypracování pevnějších statistických a informačně teoretických základů pro strojové učení a praktickým a komerčním úspěchům různých aplikací UI na specifické úlohy nebo oblasti – výzkum UI získal zpět něco ze své ztracené prestiže. Jeho dřívější dějiny však na tomto oboru možná zanechaly jisté kulturní následky, jejichž vlivem se mnozí zavedení badatelé nyní zdráhají projevovat přespříliš velké ambice. Nils Nilsson, jeden z veteránů oboru, si stěžuje, že jeho dnešní kolegové postrádají onu myšlenkovou odvahu, která poháněla vpřed průkopníky jeho generace:

> Myslím si, že starost o „serióznost" působí na některé badatele v UI dusivě. Slýchávám je říkat věci jako: „UI bývala kritizována pro svou nabubřelost. Když jsme teď učinili solidní pokroky, neměli bychom riskovat, že opět přestaneme být brány vážně." Jedním z výsledků tohoto konzervativismu je, že se zvýšená pozornost věnuje „slabé UI" – tomu druhu UI, který má poskytovat pomůcky pro lidské myšlení –, a že se naopak zájem odvrací od „silné UI" – té, která se pokouší mechanizovat inteligenci lidské úrovně.[75]

Podobný názor vyjádřilo několik dalších zakladatelských osobností, mezi nimi Marvin Minsky, John McCarthy a Patrick Winston.[76]

V posledních letech došlo k oživení zájmu o UI, což by mohlo vést i k obnově snah zaměřených na *obecnou* umělou inteligenci (tj. na to, co Nilsson nazývá „silnou UI"). Dnešní projekty by kromě rychlejšího hardwaru těžily rovněž z velkého rozmachu řady podoborů UI, softwarového inženýrství obecně a také sousedních disciplín, jako je výpočetní neurověda. Známkou toho, že se zvětšuje poptávka po kvalitních informacích a osvětě v oboru, je reakce na nabídku úvodního bezplatného online kurzu umělé inteligence uspořádaného na podzim roku 2011 Sebastianem Thrunem a Peterem Norvigem na Stanfordově univerzitě. Zapsalo se do něj asi 160 000 studentů z celého světa (a 23 000 z nich jej dokončilo).[77]

Názory expertů na budoucnost UI jsou nesmírně rozmanité. Neshody se týkají toho, jakých forem by UI mohla nakonec nabýt, i toho, kdy by těchto forem nabyla. Slovy jedné nedávné studie jsou předpovědi o budoucím vývoji umělé inteligence „stejně sebejisté, jako jsou vzájemně odlišné".[78]

Současné rozložení názorů nebylo pečlivě prozkoumáno, ale hrubou představu o něm lze získat z různých menších průzkumů a neformálních pozorování. Zvláště se můžeme opřít o řadu nedávných průzkumů, které se členů několika relevantních expertních komunit dotazovaly, kdy očekávají vznik „strojové inteligence lidské úrovně" (SILU), definované jako „inteligence, která dokáže většinu lidských povolání vykonávat přinejmenším stejně dobře jako běžný člověk."[79] Výsledky ukazuje tabulka 2. Dohromady dávají průzkumy následující (mediánový) odhad: 10% pravděpodobnost SILU v roce 2022, 50% v roce 2040 a 90% v roce 2075. (Respondenti ve svých odhadech měli pracovat s předpokladem, že „lidská vědecká činnost bude pokračovat bez významného narušení".)

Tato čísla bychom měli brát s jistou rezervou: vzorky dotazovaných expertů jsou docela malé a nemusejí být reprezentativní. Shodují se nicméně s výsledky jiných průzkumů.[80]

Stejně tak jsou tyto výsledky v souladu s nedávno vydanými rozhovory s asi půltuctem badatelů v oborech souvisejících s UI. Například Nils Nilsson během své dlouhé a plodné kariéry pracoval na problémech prohledávání, plánování, robotiky a reprezentace znalostí, je autorem učebnic umělé inteligence a nedávno dokončil zatím nejúplnější dějiny tohoto oboru.[81] Na otázku ohledně doby příchodu SILU odpověděl následovně:[82]

pravděpodobnost 10 %: 2030
pravděpodobnost 50 %: 2050
pravděpodobnost 90 %: 2100

Tabulka 2 ***Kdy bude s udanou pravděpodobností dosaženo strojové inteligence lidské úrovně?***[83]

	10 %	50 %	90 %
FT-UI	2023	2048	2080
OUI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Medián	2022	2040	2075

Soudě podle publikovaných rozhovorů odpovídá pravděpodobnostní rozdělení uvedené profesorem Nilssonem mínění mnoha expertů v oboru – ač znovu musíme zdůraznit, že rozptyl názorů je velký: někteří odborníci jsou mnohem větší nadšenci a s jistotou příchod SILU očekávají mezi lety 2020 a 2040, zatímco jiní badatelé si jsou zase jisti, že k němu buď nedojde nikdy, nebo že je v nedohlednu.[84] Jiní dotazovaní experti se nadto domnívají, že pojem „lidská úroveň" umělé inteligence je špatně definovaný či zavádějící, případně se z jiných důvodů zdráhají oficiálně vyslovit nějakou kvantifikovanou předpověď.

Můj názor je takový, že mediánová čísla ze zmíněných průzkumů přidělují příliš malou pravděpodobnost pozdějším datům. Desetiprocentní pravděpodobnost, že SILU nebude vyvinuta před rokem 2075, nebo dokonce 2100 (pod podmínkou, že „lidská vědecká činnost bude pokračovat bez významného narušení"), se mi zdá být příliš vysoká.

Experti na UI v minulosti nebyli příliš úspěšní v předpovídání, jak rychle budou pokroky v jejich oboru probíhat a jakou budou mít podobu. Na jedné straně se některé úlohy, například hraní šachu, ukázaly být řešitelné s pomocí překvapivě jednoduchých programů; rovněž názory pochybovačů, podle kterých stroje „nikdy" nedokážou to či ono, byly mnohokrát vyvráceny. Na druhé straně odborníci (a tento omyl je častější) podceňovali, jak obtížně půjde dosáhnout toho, aby se UI dokázala efektivně vyrovnat s úkoly z praxe, a přeceňovali přednosti svých oblíbených projektů nebo technik.

Průzkum položil také dvě další otázky, které jsou pro nás relevantní.

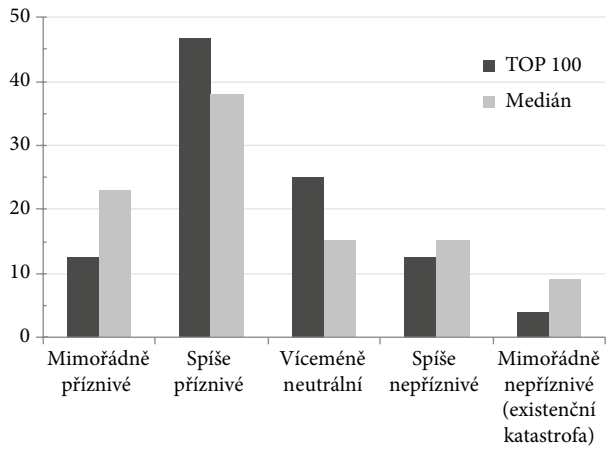
Zaprvé se respondentů ptal, jak dlouho bude ještě trvat dosažení superinteligence, jakmile již vznikne strojová inteligence lidské úrovně. Výsledky ukazuje tabulka 3.

Další otázka zněla, jaké celkové dlouhodobé dopady na lidstvo by dosažení strojové inteligence lidské úrovně mělo. Odpovědi shrnuje obrázek 2.

Mé vlastní mínění se opět od názorů vyjádřených v průzkumu trochu liší. Připisuji vyšší pravděpodobnost možnosti, že superinteligence bude vytvořena poměrně brzy po strojové inteligenci lidské úrovně. Kromě toho je můj pohled na možné následky více vyhraněný: domnívám se, že mimořádně příznivý a mimořádně nepříznivý výsledek jsou o něco pravděpodobnější než výsledek spíše neutrální. Důvody pro tento názor budou zřejmé z dalších částí knihy.

Tabulka 3 ***Za jak dlouho po SILU vznikne superinteligence?***

	Do 2 let	**Do 30 let**
TOP 100 (pravděpodobnost)	5 %	50 %
Medián (pravděpodobnost)	10 %	75 %



Obrázek 2 Celkové dlouhodobé dopady SILU.[85]

Vzhledem k malé velikosti vzorků, k výběrovému zkreslení a především k inherentní nespolehlivosti zjišťovaných subjektivních názorů bychom neměli těmto průzkumům a rozhovorům přikládat příliš velkou váhu. Nelze z nich vyvodit žádné silné závěry. Hovoří však pro jisté slabé závěry. Naznačují, že je snad rozumné předpokládat (alespoň dokud nebudou dostupná lepší data nebo analýzy), že by strojová inteligence lidské úrovně mohla s docela velkou pravděpodobností být vyvinuta někdy před polovinou 21. století a že by s pravděpodobností, jež není zcela mizivá, mohla vzniknout výrazně dříve nebo později; že by ji poměrně brzy mohla následovat superinteligence; a že může se slušnou pravděpodobností dojít k celé řadě následků včetně těch mimořádně dobrých i těch, které by znamenaly vyhynutí lidstva.[86] Přinejmenším naznačují, že toto téma stojí za bližší prozkoumání.

kapitola 2

Cesty k superinteligenci

V současnosti je obecná inteligence strojů mnohem nižší než ta lidská. A přesto se, jak tvrdíme, stroje jednoho dne stanou superinteligentními. Jak do tohoto stavu dospějeme? V této kapitole prozkoumáme několik technologií, které by nás mohly k superinteligenci dovést. Podíváme se na umělou inteligenci, úplnou emulaci mozku, biologickou kognici, na rozhraní člověk-stroj i na sítě a organizace. U všech zhodnotíme, jak hodnověrnými cestami k superinteligenci jsou. Existence několika různých cest zvyšuje pravděpodobnost, že cíle bude možné dosáhnout alespoň po jedné z nich.

Superinteligenci můžeme provizorně definovat jako *jakýkoliv intelekt, který lidské kognitivní výkony dalece překonává prakticky ve všech relevantních oblastech*.[87] O pojmu superinteligence toho řekneme více v následující kapitole, kde jej podrobíme jakési spektrální analýze, abychom rozlišili některé možné formy superinteligence. Prozatím nám právě podaná hrubá charakteristika bude stačit. Povšimněme si, že tato definice neříká nic o tom, jak bude superinteligence implementována. Stejně tak nic neříká ani o otázce *kvalií*: zda by superinteligence měla subjektivní vědomou zkušenost, to může mít pro některé (zvláště morální) otázky velký význam, avšak my se zde primárně zaměřujeme na kauzální antecedenty a následky superinteligence, nikoliv na metafyziku mysli.[88]

Šachový program Deep Fritz podle této definice superinteligencí

není, neboť je chytrý pouze v úzce vymezené oblasti šachu. Některé druhy specificky zaměřené superinteligence by však mohly být důležité. Když budeme odkazovat na superinteligentní výkony týkající se pouze nějaké dílčí oblasti, explicitně toto omezení vyznačíme. Například jako „inženýrskou superinteligenci" bychom označovali intelekt, který nejlepší lidské mozky jasně předčí ve sféře inženýrství. Není-li uvedeno jinak, užíváme tohoto termínu pouze pro systémy nadané nadlidskou *obecnou* inteligencí.

Jak bychom však mohli superinteligenci stvořit? Prozkoumejme několik možných cest.

Umělá inteligence

Čtenář by neměl očekávat, že v této kapitole nalezne návod, jak naprogramovat obecnou umělou inteligenci. Žádný takový návod zatím samozřejmě neexistuje. A kdybych jej měl k dispozici, zcela jistě bych jej nezveřejnil v knize. (Není-li ihned zřejmé proč, bude to jasné z argumentů v následujících kapitolách.)

Můžeme nicméně rozeznat některé obecné charakteristiky, které by požadovaný druh systému musel mít. Zdá se nyní zjevné, že schopnost učit se by byla nedílnou součástí samotného jádra systému určeného k dosažení obecné inteligence, nikoliv něčím, co by k němu bylo dodatečně připojeno jako rozšíření nebo doplněk. Totéž platí pro schopnost efektivně pracovat s neurčitostí a pravděpodobnostními informacemi. Dále mezi základní vlastnosti moderní UI, jež by mohla dosáhnout obecné inteligence, pravděpodobně patří schopnost získávat užitečné pojmy ze smyslových dat a svých vlastních vnitřních stavů a vkládat takto získané pojmy do flexibilních kombinatorických reprezentací, aby mohly být využity v logickém a intuitivním usuzování.

„Stará dobrá umělá inteligence" se na učení, neurčitost ani utváření pojmů většinou nezaměřovala, možná proto, že techniky pro zvládání těchto aspektů inteligence v její době nebyly příliš rozvinuty. To však neznamená, že by myšlenky stojící v základu moderních

metod byly kdovíjak nové. Myšlenku, podle které by se jednodušší program mohl pomocí učení samostatně pozvednout k inteligenci lidské úrovně, lze nalézt nejpozději v Turingově textu o „dětském stroji" z roku 1950:

Proč se snažit o vytvoření stroje, který by simuloval mysl dospělého člověka? Neměli bychom se spíše pokusit vytvořit stroj simulující mysl dítěte? Pokud bychom tento stroj posléze náležitě vzdělávali, získali bychom tak dospělý mozek.[89]

Turing si představoval, že takový dětský stroj by mohl být vyvinut prostřednictvím iterativního procesu:

Nemůžeme očekávat, že dobrý dětský stroj najdeme na první pokus. Musíme experimentovat s výukou jednoho takového stroje a pozorovat, jak dobře se učí. Pak můžeme vyzkoušet další a zjistit, jestli je lepší, nebo horší. Zjevně existuje spojitost mezi tímto procesem a evolucí… Můžeme však doufat, že tento proces bude časově úspornější než evoluce. Přežití nejzdatnějších je pomalou metodou pro měření výhod. S uplatněním své inteligence by experimentátor měl být schopen proces urychlit. Stejně důležitá je skutečnost, že si experimentátor nebude muset vystačit s nahodilými mutacemi. Pokud dokáže vystopovat příčinu nějaké slabiny, bude pravděpodobně schopen vymyslet takový druh mutace, který tuto slabinu odstraní.[90]

Víme, že slepé evoluční procesy mohou vytvořit obecnou inteligenci lidské úrovně, poněvadž tak již přinejmenším jednou učinily. Evoluční procesy obdařené předvídavostí – tj. genetické programy navržené a řízené inteligentním lidským programátorem – by měly být schopny podobného výsledku dosáhnout mnohem efektivněji. Na základě této úvahy se někteří filozofové a vědci, mezi nimi David Chalmers a Hans Moravec, snaží dokázat, že UI lidské úrovně je nejen teoretickou možností, nýbrž že je realizovatelná v průběhu tohoto století.[91] Jejich základní myšlenka je taková, že evoluci a lidskou technologii můžeme vzájemně porovnat s ohledem na jejich

schopnost vytvořit inteligenci, čímž zjistíme, že lidská technologie již nyní v některých oblastech evoluci převyšuje a zanedlouho ji pravděpodobně předežene i v těch ostatních. Skutečnost, že evoluce vytvořila inteligenci, tudíž naznačuje, že lidská technologie toho bude brzy schopna též. Moravec proto (už v roce 1976) napsal:

> Existence několika příkladů inteligence, které vznikly za těchto omezení, by nám měla dodat velkou míru jistoty, že zakrátko dokážeme dosáhnout téhož. Analogii této situace představují dějiny letadel těžších než vzduch. Než se naší kultuře podařilo tyto stroje vytvořit, ukazovali nám ptáci, netopýři a hmyz jasně, že to je možné.[92]

Při vyvozování závěrů z takovýchto úvah bychom však měli být opatrní. Je pravda, že evoluce zplodila létající organismy těžší než vzduch a že se následně i lidským inženýrům podařilo sestrojit letadla těžší vzduchu (byť pomocí velmi odlišného mechanismu). Daly by se uvést i jiné příklady, jako jsou sonar, magnetická navigace, chemické zbraně, fotoreceptory a nejrůznější mechanické a kinetické schopnosti. Stejně tak bychom však mohli poukázat na oblasti, v nichž se lidští inženýři zatím evoluci vyrovnat nedokázali: pokud jde například o morfogenezi, sebeopravování či imunitní obranu, zaostávají lidské pokusy daleko za úspěchy přírody. Moravcovy argumenty nám tudíž nemohou dodat „velkou míru jistoty", že k umělé inteligenci lidské úrovně dospějeme „zakrátko". Evoluce inteligentního života přinejlepším stanovuje jistou horní hranici pro to, jak může být stvoření inteligence obtížné. Avšak tato horní hranice by mohla být umístěna mnohem výše, než kam sahají současné lidské schopnosti.

Jinou variantou evolučního argumentu pro uskutečnitelnost UI je myšlenka, podle které bychom mohli dosáhnout výsledků srovnatelných s výsledky biologické evoluce, pokud bychom nechali genetické algoritmy běžet na dostatečně rychlých počítačích. Tato verze argumentu tedy předkládá konkrétní metodu, jejímž prostřednictvím by inteligence mohla být vytvořena.

Skutečně však budeme brzy mít k dispozici výpočetní výkon

postačující k tomu, abychom dokázali zopakovat příslušné evoluční procesy, které umožnily zrod lidské inteligence? Odpověď záleží na tom, o kolik v příštích dekádách výpočetní technologie pokročí, i na tom, jak velký výpočetní výkon by byl zapotřebí ke spuštění genetických algoritmů se stejnou schopností optimalizace, jakou měl evoluční proces v dějinách našeho druhu. Přestože nás sledování této argumentační linie nakonec dovede k výsledku, jenž bude neuspokojivým způsobem neurčitý, je poučné pokusit se o hrubý odhad (viz rámeček 3). Když nic jiného, obrátí toto cvičení naši pozornost k některým zajímavým neznámým faktorům.

Závěr těchto úvah je takový, že výpočetní zdroje potřebné k prostému zopakování relevantních evolučních procesů, které na Zemi vedly ke vzniku lidské inteligence, jsou daleko mimo náš dosah – a mimo náš dosah by zůstaly, i pokud by po celé jedno století vývoj počítačů pokračoval dle Mooreova zákona (srovnej obrázek 3). Můžeme však hodnověrně tvrdit, že když s použitím několika očividných zlepšení přírodního výběru nastavíme prohledávací proces tak, aby vznik inteligence měl za svůj cíl, dosáhneme ve srovnání s mechanickým napodobením přírodních evolučních procesů mnohem větší efektivity. Je však velmi obtížné změřit, jak velké by tyto dosažitelné nárůsty efektivity byly. Nedokážeme ani říci, zda by odpovídaly pěti, nebo dvaceti pěti řádům. Bez dalšího rozpracování proto evoluční argumenty nemohou smysluplně vytyčit hranice pro naše očekávání týkající se toho, jak obtížně lze strojovou inteligenci lidské úrovně vytvořit nebo v jakém časovém horizontu by se to mohlo podařit.

Takovéto evoluční úvahy obsahují ještě jeden problém, kvůli němuž je obtížné na jejich základě stanovit (třeba jen velmi volně) horní hranici pro obtížnost vzniku inteligence. Bylo by chybou soudit, že vyvinul-li se inteligentní život na planetě Zemi, existovala vysoká apriorní pravděpodobnost, že příslušné evoluční procesy ke vzniku inteligence povedou. Tento úsudek je nesprávný, neboť nebere v úvahu efekt výběru pozorovatele, jenž zaručuje, že všichni pozorovatelé budou vždy pocházet z planety, na níž inteligentní život vznikl, ať už bylo jakkoliv pravděpodobné nebo nepravděpodobné, že k tomu

Rámeček 3 Čeho by bylo zapotřebí k zopakování evoluce?

Ne všechno z toho, čeho evoluce dosáhla v průběhu vývoje lidské inteligence, má relevanci pro lidského inženýra, který se snaží prostřednictvím umělé evoluce vyvinout strojovou inteligenci. Jenom malá část evoluční selekce na Zemi byla selekcí inteligence. Konkrétněji řečeno, možná jenom velmi malá část celkového evolučního výběru se týkala problémů, které lidští inženýři nemohou snadno obejít. Protože například své počítače můžeme pohánět elektrickou energií, nemusíme kvůli vytvoření inteligentních strojů znovu vynalézat molekuly zapojené do buněčné energetické ekonomiky – taková molekulární evoluce metabolických drah však možná vyčerpala velkou část z celkové selekční síly, kterou měla evoluce během dějin naší planety k dispozici.[93]

Dalo by se tvrdit, že vhledy klíčové pro UI jsou vtěleny do struktury nervových soustav, které vznikly před méně než miliardou let.[94] Budeme-li zastávat tento názor, bude tím počet relevantních „experimentů" dostupných evoluci drasticky omezen. Na světě dnes existuje asi $4–6 \times 10^{30}$ prokaryot, ale jenom 10^{19} zástupců hmyzu a méně než 10^{10} lidí (přičemž velikost populace před neolitickou revolucí byla o několik řádů nižší).[95] Tato čísla nejsou až tak hrozivá.

Evoluční algoritmy však potřebují nejen varianty, mezi nimiž vybírají, ale také kriteriální („*fitness*") funkci, která jednotlivé varianty ohodnocuje, a ta je obvykle jejich výpočetně nejnákladnější složkou. Kriteriální funkce pro evoluci umělé inteligence bude pravděpodobně ke zhodnocení zdatnosti potřebovat simulaci nervového vývoje, učení a kognice. Proto snad uděláme lépe, když nebudeme věnovat pozornost jednoduše celkovému množství organismů s komplexními nervovými soustavami, nýbrž spíše počtu neuronů v těch biologických organismech, jejichž simulace by mohla být k napodobení evoluční kriteriální funkce zapotřebí. Hrubý odhad tohoto čísla můžeme učinit, když se zaměříme na hmyz, jenž vévodí pozemské zvířecí biomase (jenom mravenci tvoří podle odhadů jejích 15–20 %).[96] Mezi velikostmi hmyzích mozků panují značné rozdíly, přičemž velké a sociální druhy hmyzu jsou vybaveny většími mozky. Včelí mozek má těsně pod 10^6 neuronů, mozek

octomilky 10^5 neuronů a mozek mravence se nachází se svými 250 000 neurony mezi nimi.[97] Většina zástupců malého hmyzu má mozek o velikosti pouhých několika tisíc neuronů. Provedeme-li z opatrnosti raději vyšší odhad a přisoudíme-li všem 10^{19} zástupcům hmyzu stejný počet neuronů, jako má moucha, dospějeme k celkovému počtu 10^{24} hmyzích neuronů. Započítáme-li také vodní klanonožce, ptáky, plazy, savce atd., zvětší se toto číslo o jeden řád na 10^{25}. (Naproti tomu lidí v době před neolitickou revolucí žilo jenom 10^7, přičemž každý z nich měl pod 10^{11} neuronů: celkově tedy existovalo méně než 10^{18} lidských neuronů. U lidí ovšem neurony mají větší množství synapsí.)

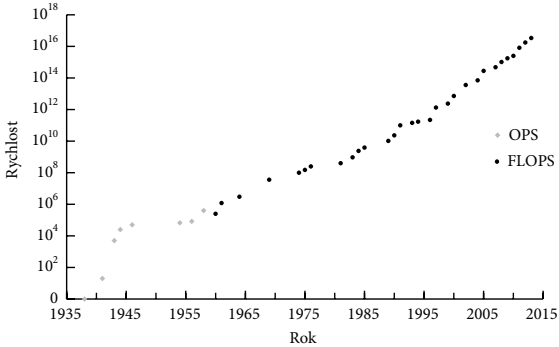
Výpočetní náročnost simulace jednoho neuronu závisí na tom, o jak detailní simulaci jde. Extrémně jednoduché neuronové modely používají k simulaci jednoho neuronu (v reálném čase) asi 1000 operací v plovoucí řádové čárce za sekundu (*floating operations per second* – FLOPS). Elektrofyziologicky realistický Hodgkinův-Huxleyho model používá 1 200 000 FLOPS. U podrobnějšího multikompartmentálního modelu by toto číslo bylo o tři až čtyři řády vyšší, zatímco vysokoúrovňové modely, které abstraktně simulují systémy neuronů, by oproti jednodušším modelům potřebovaly o dva až tři řády FLOPS méně.[98] Kdybychom měli simulovat 10^{25} neuronů v průběhu miliardy let evoluce (což je delší doba, než po kterou existují nervové soustavy v podobě, v jaké je známe) a dovolili bychom svým počítačům běžet po jeden rok, potřebovali bychom něco mezi 10^{31} a 10^{44} FLOPS. Pro srovnání, čínský počítač Tianhe-2, k září 2013 nejvýkonnější superpočítač na světě, dosahuje pouze $3{,}39 \times 10^{16}$ FLOPS. V posledních desetiletích se výkon běžných počítačů o jeden řád zvyšoval přibližně za 6,7 roku. Ani pokud by po celé století růst pokračoval podle Mooreova zákona, nepodařilo by se nám zmíněný rozdíl ve výkonnosti překonat. Specializovanější hardware nebo delší doba běhu počítačů by výkon zvětšily jenom o několik řádů.

Tento odhad je opatrný ještě v jednom ohledu. Evoluce k lidské inteligenci dospěla, aniž by to bylo jejím cílem. Jinak řečeno, kriteriální funkce vyhodnocující přírodní organismy nevybírají jenom inteligenci nebo její předchůdce.[99] Inteligentní organismy nemusejí být

selektovány ani v prostředích, kde lepší schopnost zpracovávat informace přináší různé výhody, jelikož zlepšení v inteligenci s sebou může nést (a často skutečně nese) velké náklady, například vyšší spotřebu energie či pomalejší dospívání, a tyto náklady mohou nad prospěchem plynoucím z chytřejšího chování převážit. Hodnotu inteligence snižují také extrémně nebezpečná prostředí: čím menší je očekávaná délka života, tím méně času je na to, aby se náklady na lepší schopnost učit se vyplatily. Kvůli zmenšenému selekčnímu tlaku na rozvoj inteligence se inovace, které inteligenci zvyšují, šíří pomaleji, a následkem toho se také zmenšuje množství příležitostí, aby evoluce vybírala další inovace, jež závisejí na těch předchozích. Evoluce může nadto zamrznout na lokálních optimech, kterých by si lidé povšimli a obešli je jiným vzájemným vyvážením exploatace a explorace nebo aplikací řady čím dál obtížnějších testů inteligence.[100] A jak již víme, evoluce mnoho své selekční síly vynakládá na znaky, které s inteligencí nesouvisejí (příkladem je „závod Červené královny"* mezi imunitními systémy a parazity). Evoluce plýtvá prostředky, když pokračuje s vytvářením mutací, které se trvale ukazují být smrtelné, a nedokáže využít statistických podobností mezi účinky různých mutací. Lidský inženýr, jenž by k rozvoji inteligentního softwaru využíval evoluční algoritmy, by se všem těmto nedostatkům přírodního výběru (chápaného jako prostředek k rozvoji inteligence) mohl poměrně snadno vyhnout.

Lze věřit tomu, že odstranění takovýchto nedostatků by požadovaný výpočetní výkon, výše odhadovaný na 10^{31} až 10^{44} FLOPS, o několik řádů snížilo. Bohužel však těžko můžeme vědět, o kolik přesně. Už jenom učinit hrubý odhad je obtížné – nemáme tušení, jestli by úspora činila pět, deset nebo dvacet pět řádů.[101]

* Jako „závod Červené královny" se označuje situace, v níž se jednotlivé druhy musejí neustále vyvíjet už proto, aby udržely krok se svým protivníkem, a tak unikly vyhynutí. Termín odkazuje ke knížce *Za zrcadlem, a s čím se tam Alenka setkala* Lewise Carrolla: jak se dozvídáme od šachové královny, s níž se Alenka setkává na začátku knížky, ve světě za zrcadlem člověk musí neustále běžet ze všech sil, jen aby zůstal na jednom místě. Zmíněná královna se v originále skutečně jmenuje „Red Queen", v českých překladech je to však Černá královna; pro onen biologický pojem se u nás nicméně vžil překlad „závod Červené královny" či „efekt Červené královny". – Pozn. překl.



Obrázek 3 Výkonnost superpočítačů. V úzkém smyslu se jako „Mooreův zákon" označuje pozorování, podle kterého se počet tranzistorů na integrovaných obvodech po několik desetiletí zdvojnásobil přibližně každý druhý rok. Často se však tento termín užívá také jako označení pro obecnější poznatek, že v oblasti výpočetních technologií podobným tempem exponenciálně rostou i jiné ukazatele výkonnosti. Na tomto grafu je maximální rychlost nejrychlejšího počítače na světě zakreslena jako funkce času (s logaritmickým měřítkem ve svislé ose). V posledních letech sériová rychlost procesorů stagnuje, ale díky rostoucímu užívání paralelizace se celkový počet provedených výpočtů stále drží trendové přímky.[102]

na jakékoliv planetě téhož typu dojde. Předpokládejme například, že ke vzniku inteligentního života bylo kromě systematických účinků přirozeného výběru zapotřebí také ohromné množství *šťastné náhody* – tak velké, že se inteligentní život vyvine jenom na jedné z 10^{30} planet, kde vzniknou jednoduché replikátory. Pokud bychom se v takovém případě snažili s pomocí svých genetických algoritmů zopakovat výsledky přírodní evoluce, zjistili bychom možná, že potřebujeme provést nějakých 10^{30} simulací, než najdeme jednu, v níž se všechno seběhne tím správným způsobem. To se zdá být zcela slučitelné s naším pozorováním, že se zde na Zemi život skutečně vyvinul. Tuto epistemologickou překážku můžeme částečně obejít jenom s pomocí opatrných a poněkud spletitých úvah – tak, že budeme analyzovat případy konvergentní evoluce znaků souvisejících s inteligencí a věnovat se delikátním otázkám teorie výběru pozorovatele. Pokud si tu práci nedáme, nebudeme moci vyloučit možnost, že skutečná horní hranice pro výpočetní nároky na zopakování evoluce by mohla být o třicet řádů (nebo o nějaké podobně velké číslo) vyšší než ta, kterou jsme odvodili v rámečku 3.[103]

Chce-li člověk argumentovat pro uskutečnitelnost umělé inteligence, může dále poukázat na existenci lidského mozku a navrhnout, aby byl použit jako předloha pro strojovou inteligenci. Můžeme rozlišit několik odlišných verzí tohoto přístupu podle toho, jak přesně chtějí funkce mozku napodobovat. Na jednom konci (odpovídajícím snaze o velmi přesnou nápodobu) stojí myšlenka *úplné emulace mozku*, které se budeme věnovat v následujícím oddílu. Na druhém konci se pak nacházejí přístupy, jež se fungováním mozku inspirují, ale nepokoušejí se o jeho nízkoúrovňovou nápodobu. Pokroky v neurovědě a kognitivní psychologii by (s přispěním lepších nástrojů) měly nakonec vést k odhalení obecných principů fungování mozku. Těmito poznatky by se pak mohly řídit snahy o vytvoření UI. Jedním příkladem UI techniky inspirované mozkem jsou neuronové sítě, s nimiž jsme se již setkali. Další ideou, kterou strojové učení převzalo od vědy o mozku, je myšlenka hierarchického uspořádání vnímání. Zkoumání zpětnovazebního učení bylo (alespoň zčásti) motivováno rolí, jakou toto učení hraje v psychologických teoriích zvířecí kognice, a techniky inspirované těmito teoriemi (například „TD-algoritmus“) jsou nyní v UI široce uplatňovány.[104] Případy tohoto druhu se v budoucnosti určitě budou hromadit. Protože existuje jenom omezené (a možná velmi malé) množství základních mechanismů fungování mozku, měla by je věda o mozku cestou postupných pokroků nakonec odhalit všechny. Než k tomu však dojde, mohl by cílové pásky dosáhnout hybridní přístup, jenž by techniky inspirované mozkem kombinoval s některými čistě umělými metodami. V takovém případě by se výsledný systém nemusel mozku zřetelně podobat, ač by při jeho vývoji byly využity některé poznatky vyvozené ze zkoumání mozku.

Skutečnost, že máme k dispozici mozek coby vzor inteligentního systému, představuje silnou oporu pro tvrzení, že strojovou inteligenci nakonec sestrojit půjde. To však neznamená, že bychom mohli předpovědět, kdy se tak stane, neboť lze obtížně odhadnout tempo budoucích objevů ve vědě o mozku. Můžeme říci jenom tolik, že čím dále do budoucnosti hledíme, tím se stává pravděpodobnějším, že tajemství fungování mozku budou dešifrována

v takové míře, abychom na základě těchto zkoumání mohli vytvořit strojovou inteligenci.

Různí lidé pracující na strojové inteligenci mají různé názory na to, jak slibné jsou neuromorfní přístupy ve srovnání s těmi zcela syntetickými. Existence ptáků prokázala, že věci těžší než vzduch mohou létat, a podnítila snahy o vytvoření létajících strojů. Přesto první fungující letadla nemávala křídly. Zatím není jisté, zda to se strojovou inteligencí bude jako s létáním, jehož lidé dosáhli s pomocí umělého mechanismu, nebo jako s objevem ohně, který jsme na počátku ovládli skrze nápodobu ohňů vyskytujících se v přírodě.

Turingova myšlenka programu, do něhož by jeho obsahy nebyly předem vloženy, nýbrž by se většině z nich naučil, může být platná pro neuromorfní i pro syntetické přístupy ke strojové inteligenci.

Variaci na Turingovu koncepci dětského stroje představuje myšlenka „zárodečné UI“.[105] Zatímco dětský stroj by podle Turingových představ zřejmě měl poměrně fixní architekturu, jež by jednoduše rozvíjela své vnitřní možnosti prostřednictvím hromadění *obsahu*, zárodečná UI by byla sofistikovanější umělou inteligencí, schopnou zlepšovat právě svou vlastní *architekturu*. V prvních stadiích vývoje zárodečné UI by k takovým zlepšením docházelo hlavně na základě metody pokusu a omylu, zisku nových informací či s přispěním programátorů. V pozdějších stadiích by však zárodečná UI měla být schopna svému vlastnímu fungování *porozumět* natolik, aby dokázala naprojektovat nové algoritmy a výpočetní struktury, jimiž si sama napomůže k lepším kognitivním výkonům. Toto potřebné porozumění by mohlo vzniknout buď tak, že zárodečná UI dosáhne dostatečné úrovně obecné inteligence napříč mnoha doménami, anebo tím způsobem, že se zlepší v nějaké obzvláště relevantní oblasti, například v informatice nebo matematice.

To nás přivádí k dalšímu důležitému konceptu, jímž je „rekurzivní sebezlepšování“. Úspěšná zárodečná UI by byla schopna iterativně vylepšovat sebe samu: raná verze této UI by dokázala naprojektovat vylepšenou verzi sebe samé, tato její vylepšená verze by – jsouc chytřejší než ta původní – možná dokázala naprojektovat ještě chytřejší verzi sebe samé (a tak dále).[106] Za určitých podmínek by takový

proces rekurzivního sebezlepšování mohl pokračovat dost dlouho, aby vyústil v inteligenční explozi – událost, kdy inteligence systému za krátkou dobu vzroste z poměrně skromné kognitivní úrovně (možná ve většině ohledů nižší než lidské, ale vysoké v oblasti programování a výzkumu v UI) na úroveň radikální superinteligence. K této důležité možnosti se vrátíme v kapitole 4, kde bude dynamika takové události analyzována podrobněji. Tento model, jak si můžeme povšimnout, naznačuje, že může dojít k překvapením: pokusy o vytvoření UI třeba dlouho budou v podstatě naprosto neúspěšné, dokud nenajdeme tu poslední, klíčovou chybějící součástku – a v ten okamžik by zárodečná UI mohla získat schopnost dlouhodobě se rekurzivně zlepšovat.

Než tento oddíl uzavřeme, měli bychom zdůraznit ještě jednu věc, totiž že UI se nemusí příliš podobat lidské mysli. Umělé inteligence by od nás mohly být zcela odlišné – a je dokonce pravděpodobné, že většina z nich taková bude. Měli bychom počítat s tím, že budou mít výrazně jinou kognitivní architekturu a – alespoň v raných fázích svého vývoje – jinou kombinaci silných a slabých stránek než biologická inteligence (jak však ukážeme později, všechny počáteční slabiny by mohly nakonec překonat). Nadto by se i jejich motivační systémy mohly zásadně odlišovat od těch lidských. Nemáme důvod očekávat, že by běžná UI byla motivována láskou, nenávistí, pýchou nebo jiným obvyklým lidským citem tohoto typu: opětovné stvoření těchto komplexních adaptivních znaků u UI by bylo nákladné a vyžadovalo by vynaložení vědomého úsilí. To představuje zároveň velký problém i velkou příležitost. K tématu motivace UI se vrátíme v pozdějších kapitolách, pro naši argumentaci však má natolik klíčový význam, že se vyplatí mít je neustále na paměti.

Úplná emulace mozku

Úplná emulace mozku (známá také jako „*uploading*“) je metoda stvoření inteligentního softwaru, jež spočívá v nasnímání a následném přesném namodelování výpočetní struktury biologického mozku.

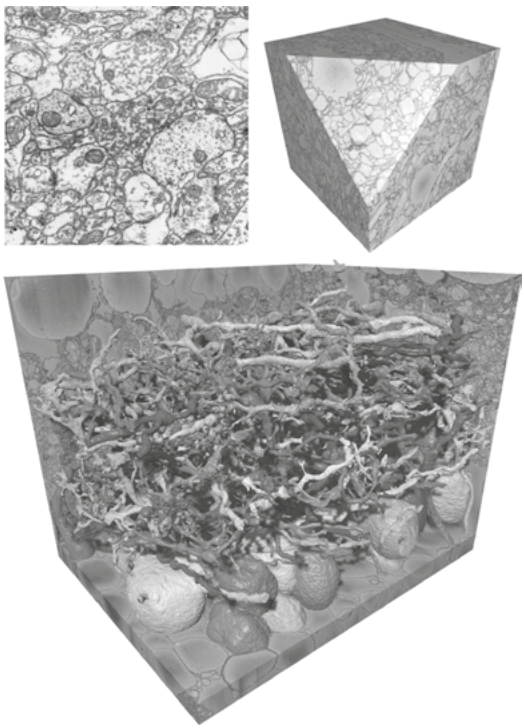
Jde tedy o krajní případ inspirace přírodou, o bezostyšné plagiátorství. K jejímu vytvoření je zapotřebí uskutečnit následující kroky.

Zaprvé, vytvořit dostatečně podrobný snímek nějakého konkrétního lidského mozku. To by mohlo obnášet posmrtnou stabilizaci mozku prostřednictvím vitrifikace (procesu, jímž se tkáň přeměňuje v jakýsi druh skla). Mozkovou tkáň by pak nějaký stroj mohl nařezat na tenké plátky, jež by následně byly předány jiným strojům (například velkému množství elektronových mikroskopů) k nasnímání. V této fázi by mohla být k odkrytí rozličných strukturních a chemických vlastností na vzorky aplikována různá barviva. Snímkování by mohlo probíhat paralelně na mnoha strojích, aby se dalo zpracovat více plátků mozku najednou.

Zadruhé, nahrát surová data ze snímkovacích přístrojů do počítače. Program pro automatické zpracování snímků pak zrekonstruuje trojrozměrnou neuronovou síť, která byla nositelem kognice v původním mozku. V praxi by oba kroky mohly probíhat souběžně, čímž by se snížilo množství snímků o vysokém rozlišení, jež musejí být uloženy ve vyrovnávací paměti. Výsledná mapa mozku se poté zkombinuje s databází neurovýpočetních modelů neuronů či různých prvků neuronů (například konkrétních druhů synaptických spojů). Obrázek 4 ukazuje, co dokážou vytvořit dnešní mikroskopy a technologie na zpracování snímků.

Ve třetí fázi by pak neurovýpočetní struktura, k níž jsme dospěli v druhém kroku, byla implementována na dostatečně výkonném počítači. Pokud by celá procedura byla zcela úspěšná, byla by jejím výsledkem digitální reprodukce původního intelektu, která by ponechala jeho paměť i osobnost netknutou. Emulovaná lidská mysl nyní existuje v podobě počítačového softwaru. Může buď obývat virtuální realitu, anebo prostřednictvím robotických údů interagovat s vnějším světem.

Pro úplnou emulaci mozku není zapotřebí, abychom přišli na to, jak funguje lidská kognice nebo jak naprogramovat umělou inteligenci. Musíme jenom rozumět nízkoúrovňovým funkčním charakteristikám základních výpočetních prvků mozku. Úspěch úplné emulace nevyžaduje žádný zásadní konceptuální ani teoretický průlom.



Obrázek 4 Rekonstrukce 3D neuroanatomie na základě snímků z elektronového mikroskopu. *Vlevo nahoře*: běžná elektronová mikrofotografie znázorňující řez nervovou hmotou – dendrity a axony. *Vpravo nahoře*: objemový obraz nervové tkáně ze sítnice králíka získaný technikou sériové blokové řádkovací mikroskopie.[107] Z jednotlivých 2D snímků byla poskládána krychle s hranou o délce asi 11 μm. *Dole*: rekonstrukce části neuronových výběžků vyplňujících objem neuropilu, jež byla vygenerována automatickým segmentačním algoritmem.[108]

Co však úplná emulace mozku vyžaduje, to jsou poměrně pokročilé technologie. Ty musí umožňovat následující akce: 1) *snímkování:* vysoce výkonné mikroskopy s dostatečným rozlišením a schopností detekovat důležité vlastnosti; 2) *převod:* automatická analýza snímků, která ze surových nasnímaných dat vytvoří trojrozměrný model relevantních neurovýpočetních prvků; 3) *simulace:* hardware natolik výkonný, aby na něm mohla být výsledná výpočetní struktura spuštěna (viz tabulku 4). (Ve srovnání s těmito náročnějšími kroky představuje konstrukce jednoduché virtuální reality či robotického

těla se vstupním audiovizuálním kanálem a nějakým prostým výstupním kanálem relativně snadný úkol. Jednoduchá, avšak na nejzákladnější úrovni postačující[109] zařízení vstup/výstup lze zřejmě vybudovat už s dnešními technologiemi.)

Tabulka 4 ***Schopnosti potřebné pro úplnou emulaci mozku***

Snímkování	Předzpracování/ fixování		Náležitá příprava mozku, uchování jeho relevantního stavu a mikrostruktury
	Fyzická manipulace		Metody manipulace s fixovanými mozky a kusy tkání před snímkováním, v jeho průběhu a po něm
	Zobrazování	Objem	Schopnost nasnímat celý objem mozku s vynaložením přiměřeného množství času a nákladů
		Rozlišení	Snímkování v dostatečném rozlišení, které dovolí mozek rekonstruovat
		Informace o funkcích	Snímkování schopné detekovat funkčně důležité vlastnosti tkáně
Převod	Zpracování obrazu	Geometrická úprava	Odstraňování zkreslení způsobených nedokonalým snímkováním
		Interpolace dat	Nahrazování chybějících dat
		Odstranění šumu	Zlepšování kvality snímku
		Trasování	Odhalení struktury tkáně a její zpracování do podoby konzistentního 3D modelu

	Interpretace snímku	Identifikace typů buněk	Identifikování typů buněk
		Identifikace synapsí	Identifikování synapsí a jejich konektivity
		Odhad parametrů	Odhadování funkčně důležitých parametrů buněk, synapsí a dalších prvků
		Ukládání do databáze	Úsporné skladování výsledného inventáře prvků
	Softwarový model nervového systému	Matematický model	Modelování prvků a jejich chování
		Efektivní implementace	Implementace modelu
Simulace	Ukládání		Skladování původního modelu a jeho současného stavu
	Přenosová rychlost		Efektivní komunikace mezi procesory
	CPU		Výkon procesorů umožňující spuštění simulace
	Simulace těla		Simulace těla, jež emulaci dovolí interagovat s virtuálním nebo (prostřednictvím robota) skutečným okolním prostředím
	Simulace prostředí		Virtuální prostředí pro virtuální tělo

Máme dobré důvody se domnívat, že potřebné technologie dosažitelné jsou, byť nikoliv v blízké budoucnosti. Slušné výpočetní modely mnoha typů neuronů a neuronových procesů existují již nyní.

Byl vyvinut software na rozpoznávání obrazu, který dokáže napříč množstvím dvourozměrných snímků trasovat axony a dendrity (ač jeho spolehlivost se ještě musí zvýšit). A existují zobrazovací nástroje, které nám poskytují dostatečné rozlišení – řádkovací tunelový mikroskop nám umožňuje „vidět" jednotlivé atomy, což je mnohem vyšší rozlišení, než potřebujeme. Jakkoliv však naše současné znalosti a schopnosti nasvědčují, že neexistuje žádná principiální překážka pro vývoj potřebných technologií, byl by k tomu, aby se úplná emulace mozku stala dosažitelnou, zjevně zapotřebí dlouhotrvající technologický pokrok.[110] Například by mikroskopy kromě potřebného rozlišení musely mít rovněž potřebnou rychlost. Zobrazování dostatečně rozměrných povrchů s pomocí řádkovacího tunelového mikroskopu, jenž má rozlišení na úrovni atomů, by bylo přespříliš pomalé na to, aby bylo proveditelné. Mnohem větší šance na úspěch by skýtal elektronový mikroskop s menším rozlišením, pro práci s ním bychom však museli mít k dispozici nové metody přípravy a barvení nervové tkáně, jež by zviditelnily potřebné detaily, třeba jemnou synaptickou strukturu. Stejně tak bychom potřebovali velké rozšíření neurovýpočetních databází i významné zlepšení automatického zpracování obrazu a interpretace mikroskopických snímků.

Obecně řečeno je úplná emulace mozku ve srovnání s umělou inteligencí založena méně na teoretickém porozumění a více na technologických schopnostech. Jak pokročilé technologie jsou pro ni zapotřebí, to záleží na stupni abstrakce, na kterém je mozek emulován. Zde existuje nepřímá úměra mezi úrovní technologií a mírou teoretického porozumění. Obecně vzato, čím horší jsou naše snímkovací zařízení a čím slabší jsou naše počítače, tím méně se můžeme spoléhat na simulaci nízkoúrovňových chemických a elektrofyziologických procesů v mozku a tím lépe musíme pochopit emulovanou výpočetní architekturu, abychom tak mohli vytvořit abstraktnější reprezentaci příslušných funkcí.[111] Kdybychom naopak měli dostatečně pokročilou snímkovací technologii a ohromně výkonné počítače, mohli bychom i s relativně omezeným porozuměním mozku k jeho emulaci dospět hrubou silou. Můžeme si představovat, že bychom – v nerealistickém krajním případě – mohli mozek

emulovat na úrovni jeho elementárních částic s pomocí kvantově mechanické Schrödingerovy rovnice. Pak bychom si vystačili výhradně s existujícími fyzikálními poznatky a obešli bychom se bez jakéhokoliv biologického modelu. Tento extrémní případ by však kladl naprosto nerealistické nároky na výpočetní kapacitu a techniky získávání dat. Mnohem spíše by mohla vzniknout emulace, která by pracovala s jednotlivými neurony a jejich maticí konektivity, jakož i s částí struktury jejich dendritických stromů a možná s některými stavovými proměnnými jednotlivých synapsí. Nesimulovali bychom zvlášť každou molekulu neurotransmiterů, nýbrž bychom nahrubo modelovali jejich kolísající koncentraci.

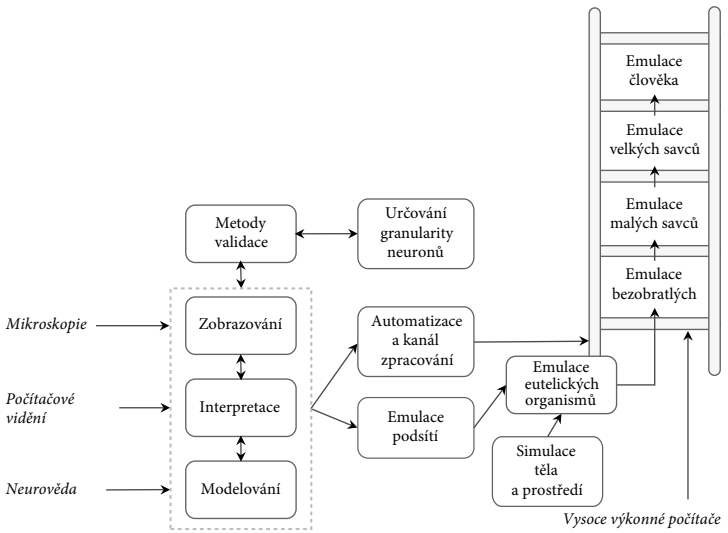
Máme-li odhadnout, zda je úplná emulace mozku realizovatelná, musíme pochopit, co je kritériem úspěchu. Cílem není vytvořit tak podrobnou a věrnou simulaci mozku, že by s její pomocí bylo možné přesně předpovědět, co by probíhalo ve skutečném mozku, pokud by byl vystaven určitému sledu stimulů. Namísto toho chceme zachytit funkční vlastnosti mozku, které mají význam pro jeho výpočetní schopnosti, v takovém rozsahu, aby výsledná emulace dokázala vykonávat intelektuální činnost. Pro tento účel jsou mnohé nepřehledné biologické detaily irelevantní.

Propracovanější analýza by rozlišila mezi různými stupni úspěšnosti emulace podle toho, kolik toho emulace uchovává ze schopností zpracovávat informace, které měl původní mozek. Například bychom mohli rozlišit: 1) *vysoce věrnou emulaci*, která má veškeré znalosti, dovednosti, schopnosti a hodnoty emulovaného mozku; 2) *zkreslenou emulaci*, jejíž dispozice jsou v některých ohledech výrazně ne-lidské, ale která z většiny dokáže vykonat tutéž intelektuální práci jako emulovaný mozek; a 3) *generickou emulaci* (jež může zároveň být zkreslená), která je trochu jako malé dítě: postrádá dovednosti a vzpomínky, které získal emulovaný dospělý mozek, avšak dokáže se naučit většinu toho, co normální člověk.[112]

Ačkoliv se zdá, že vysoce věrná emulace nakonec realizovatelná bude, zřejmě je docela pravděpodobné, že pokud bychom se touto cestou vydali, naše *první* úplná emulace mozku by na této úrovni nebyla. Než by se nám podařilo, aby emulace fungovala dokonale,

pravděpodobně bychom dosáhli toho, že by fungovala nedokonale. Také je možné, že úsilí zaměřené na emulační technologie by vedlo k vytvoření jakési neuromorfní UI, která by si osvojila některé neurovýpočetní principy odhalené během práce na emulaci a zkombinovala je se syntetickými metodami. K tomu by přitom mohlo dojít před dokončením plně funkční emulace mozku. Jak uvidíme v jedné z pozdějších kapitol, možnost takovýchto vedlejších účinků v oblasti neuromorfní UI komplikuje naše strategické úvahy, jak žádoucí by bylo vývoj emulace urychlit.

Jak daleko dnes máme k vytvoření úplné emulace mozku? Jedna nedávná studie předložila technologický plán a dospěla k závěru, že by potřebné technologie mohly být dostupné asi v polovině století (interval neurčitosti je však velký).[113] Obrázek 5 znázorňuje hlavní milníky na cestě k emulaci. Zdánlivá jednoduchost této mapy však může být zavádějící; neměli bychom podcenit, kolik práce ještě zbývá udělat. Žádný mozek dosud emulován nebyl.



Obrázek 5 Technologický plán pro úplnou emulaci mozku. Schéma vstupů, činností a milníků.[114]

Vezměme si jednoduchý modelový organismus *Caenorhabditis elegans* (háďátko obecné), průhlednou hlístici dlouhou asi 1 mm a mající 302 neuronů. Úplná matice konektivity těchto neuronů je známa od poloviny osmdesátých let 20. století, kdy byla namáhavě zmapována pomocí řezání, elektronového mikroskopování a ručního značení vzorků.[115] Nestačí ale vědět, které neurony jsou navzájem propojeny. K vytvoření emulace mozku bychom potřebovali vědět také to, které synapse jsou excitační a které inhibiční, jaká je síla spojení mezi neurony a jaké jsou různé dynamické vlastnosti axonů, synapsí a dendritických stromů. Tyto informace zatím neznáme ani v případě malé nervové soustavy *C. elegans* (byť její zjištění je nyní zřejmě v možnostech středně velkého cíleného výzkumného projektu).[116] Úspěšná emulace maličkého mozku, například mozku *C. elegans*, by nám poskytla lepší obraz, čeho by bylo zapotřebí k emulaci větších mozků.

V jistém momentu technologického vývoje – až budeme mít k dispozici techniky pro automatickou emulaci malého množství mozkové tkáně – se celý problém zredukuje na problém měřítka. Povšimněme si „žebříku" v pravé části obrázku 5. Tato vzestupná řada rámečků představuje posloupnost závěrečných vývojových stupňů, jež bychom mohli začít zdolávat, poté co se vyrovnáme s počátečními překážkami. Její jednotlivá stadia odpovídají úplné emulaci mozků modelových organismů, které mají čím dál složitější nervovou soustavu. Tato řada by mohla vypadat například takto: *C. elegans* → *včela* → *myš* → *makak rhesus* → *člověk*. Protože mezery mezi těmito příčkami jsou – přinejmenším od druhého kroku – převážně kvantitativní a zapříčiněné hlavně (byť nikoliv výhradně) rozdílnou velikostí mozku daných živočichů, mělo by být možné je překonat s pomocí relativně přímočarého rozšíření snímkovací a simulační kapacity.[117]

Jakmile začneme stoupat po tomto posledním žebříku, začne být lépe předvídatelné, kdy konečně dospějeme k úplné emulaci lidského mozku.[118] Můžeme tedy očekávat, že na cestě úplné emulace budeme o blížící se strojové inteligenci lidské úrovně předem informováni – alespoň pokud onou technologií, která dostatečně dozraje jako poslední, bude buď vysoce výkonné snímkování, nebo

výpočetní technologie schopná simulovat mozek v reálném čase. Pokud však poslední chybějící technologií bude neurovýpočetní modelování, mohl by přechod od chabých prototypů k fungující emulaci lidského mozku být náhlejší. Lze si představit scénář, v němž by se nám vzdor velkému množství nasnímaných dat a rychlým počítačům stále nedařilo neuronové modely úspěšně uvést do chodu. Když pak konečně opravíme i tu poslední závadu, může se stát, že doposud zcela dysfunkční systém – snad analogický nevědomému mozku, jenž prodělává těžký epileptický záchvat – náhle přejde do lucidního, bdělého stavu. V takovém případě by předzvěstí klíčového objevu nebyla řada fungujících emulací stále větších zvířecích mozků (vyvolávající řadu novinových titulků, jejichž velikost bude vykazovat tutéž vzestupnou tendenci). Dokonce i pro ty, kdo by této problematice věnovali pozornost, by mohlo být obtížné předem uhodnout – a to i v samotný předvečer rozhodujícího průlomu –, kolik nedostatků neurovýpočetní modely v daný okamžik dosud mají a jak dlouho bude trvat, než se je podaří opravit. (Až by úplné emulace lidského mozku dosaženo bylo, nastal by další potenciálně výbušný vývoj; této otázce se však budeme věnovat až ve 4. kapitole.)

I kdyby veškerý relevantní výzkum byl provozován veřejně, jsou tedy v případě úplné emulace mozku představitelné scénáře, které by zahrnovaly prvek překvapení. Přesto je u úplné emulace mozku ve srovnání s umělou inteligencí pravděpodobnější, že její vznik budou ohlašovat jasná znamení, neboť je závislejší na konkrétních pozorovatelných technologiích a není cele založena na teoretickém porozumění inteligenci. O této výzkumné cestě také s větší jistotou než o UI můžeme tvrdit, že v blízké budoucnosti (řekněme v následujících patnácti letech) nedosáhne svého cíle, poněvadž víme, že ještě nebylo vyvinuto několik náročných technologií, kterých by k tomu bylo zapotřebí. Oproti tomu se zdá pravděpodobné, že by se někdo *v principu* mohl prostě posadit k dnešnímu osobnímu počítači a naprogramovat zárodečnou UI; a je myslitelné, byť nepravděpodobné, že v blízké budoucnosti někdo někde přijde na to, jak to udělat.

Biologická kognice

Třetí cesta, jak dosáhnout vyšší inteligence, než jakou v současnosti disponují lidé, spočívá ve vylepšování biologických mozků. To by se v principu dalo uskutečnit bez jakýchkoliv technologií, jen prostřednictvím šlechtění. Jakýkoliv pokus realizovat v širokém měřítku klasický eugenický program by však narazil na zásadní politické a morální překážky. Pokud by nadto selekce nebyla extrémně silná, museli bychom na výraznější výsledky čekat mnoho generací. Dlouho předtím, než by taková iniciativa přinesla své plody, nám vývoj biotechnologií umožní kontrolovat lidskou genetiku mnohem přímějším způsobem a učiní jakýkoliv program lidského šlechtění zbytečným. Soustředíme se proto na metody, jež mají potenciál přinést výsledky rychleji – během několika generací nebo ještě dříve.

Kognitivní schopnosti jednotlivců lze vylepšovat řadou způsobů včetně takových tradičních metod, jako jsou vzdělávání a trénink. Vývoj nervové soustavy můžeme podporovat různými technologicky nenáročnými zásahy, například tak, že optimalizujeme výživu matky a dítěte, odstraníme z životního prostředí olovo a jiné škodliviny, vymýtíme parazity, budeme předcházet onemocněním mozku a zajistíme, abychom měli dostatek spánku a pohybu.[119] Každým z těchto prostředků bezpochyby lze kognitivní schopnosti zlepšit, ač tyto zisky pravděpodobně budou nevelké, zvláště u obyvatelstva, které už poměrně dobře živené a vzdělávané je. Těmito cestami nepochybně nedospějeme k superinteligenci, mohou však být nápomocné na okrajích inteligenčního spektra, hlavně tím, že pomohou strádajícím a umožní lépe podchytit talentované jedince na celém světě. (V mnoha chudých vnitrozemských oblastech je stále rozšířeno celoživotní snížení inteligence způsobené nedostatkem jodu – což je ostudné, neboť by tomu šlo s malými náklady, za několik centů na osobu a rok, zabránit obohacováním kuchyňské soli.)[120]

Větší zlepšení by mohla poskytnout biomedicína. Již nyní existují přípravky, které alespoň u některých testovacích subjektů údajně zlepšují paměť i schopnost soustředění a zvyšují mentální energii.[121]

(Práce na této knize byla poháněna kávou a nikotinovými žvýkačkami.) Účinnost dnešní generace „chytrých drog" je sice proměnlivá, nepatrná a vesměs pochybná, avšak v budoucnosti by přínos nootropik mohl být zřejmější a jejich vedlejší účinky menší.[122] Z neurologických i evolučních důvodů však působí nevěrohodně, že bychom mohli dramaticky zvýšit inteligenci zdravého člověka tím, že do jeho mozku dopravíme nějakou chemikálii.[123] Kognitivní fungování lidského mozku závisí na přesné souhře mnoha faktorů, která se utváří především během rozhodujících stadií embryonálního vývoje – a je velmi pravděpodobné, že k zlepšení této sebeorganizující se struktury je zapotřebí pečlivě ji vyladit, uvést do rovnováhy a zkultivovat, nikoliv do ní prostě zvnějšku nalít nějaký lektvar.

Práce s genetikou nám poskytne účinnější nástroje než psychofarmakologie. Vraťme se k myšlence genetické selekce: namísto toho, abychom se snažili uskutečnit eugenický program prostřednictvím kontroly pohlavních styků, můžeme provozovat selekci na úrovni embryí nebo pohlavních buněk.[124] Už nyní se během oplodnění ve zkumavce (*in vitro* fertilizace, IVF) využívá preimplantační genetická diagnostika, aby se u vytvořeného embrya odhalila přítomnost monogenních chorob (jako je Huntingtonova nemoc) a predispozice k některým chorobám s pozdním nástupem, například k rakovině prsu. Rovněž se jí využívá k výběru pohlaví dítěte a k zajištění shody typu jeho lidského leukocytárního antigenu s typem přítomným u jeho nemocného sourozence, jemuž po narození druhého dítěte mohou být transplantovány kmenové buňky z pupečníkové krve.[125] Množství znaků, v jejichž prospěch nebo neprospěch bude možné provádět selekci, se během jednoho nebo dvou desetiletí výrazně zvýší. Významným hybatelem pokroku v behaviorální genetice jsou klesající ceny genotypizace a genového sekvenování. Celogenomová analýza komplexních znaků, využívající studie s velkým množstvím testovacích subjektů, začíná být dostupná právě nyní a značně zlepší naše znalosti genetické architektury lidských kognitivních a behaviorálních znaků.[126] Každý znak s nezanedbatelnou mírou dědivosti včetně kognitivních schopností by pak mohl být přístupný

selekci.[127]* K selekci embryí nepotřebujeme do hloubky porozumět kauzálním drahám, jimiž geny – ve složité souhře s prostředím – produkují fenotypy: potřebujeme pouze data (velké množství dat) o genetických korelátech těch znaků, o něž nám jde.

Lze spočítat hrubé odhady toho, jak velkých zisků lze dosáhnout v různých selekčních scénářích.[128] Tabulka 5 ukazuje očekávaný nárůst inteligence, k němuž by vedla selekce z různého množství embryí za předpokladu, že bychom měli úplné informace o běžných aditivních genetických variantách, jež stojí za dědivostí inteligence (resp. její dědivostí v užším smyslu).

Tabulka 5 ***Maximální nárůst IQ při selekci ze skupiny embryí***[129]

Selekce	Nárůst bodů IQ
1 z 2	4,2
1 z 10	11,5
1 ze 100	18,8
1 z 1000	24,3
1 z 10 v pěti generacích	< 65 (kvůli klesajícím výnosům)
1 z 10 v deseti generacích	< 130 (kvůli klesajícím výnosům)
Kumulativní hranice (aditivní varianty jsou optimalizovány pro kognici)	100 + (< 300, kvůli klesajícím výnosům)

(S pouze částečnými informacemi by se efektivita selekce snížila, ale ne až tak, jak bychom mohli naivně očekávat.)[130] Není překvapivé, že při výběru z většího množství embryí dosáhneme větších zisků. Čelíme však strmě klesajícím výnosům: při výběru ze 100 embryí zisky ani zdaleka nejsou padesátkrát větší než při výběru z dvou embryí.[131]

Zajímavé je, že když se selekce rozloží na více generací, pokles výnosů se tím do značné míry utlumí. Budeme-li vybírat jedno nejlepší

* Dědivost (heritabilita) udává, do jaké míry jsou individuální rozdíly v dané vlastnosti (například v inteligenci) způsobeny genetickými faktory. Blíží-li se dědivost 0, je většina proměnlivosti znaku způsobena prostředím; blíží-li se 1, je většina variability dána genetickou výbavou. – Pozn. překl.

embryo z 10 opakovaně po 10 generací (přičemž každá nová generace bude sestávat z potomků těch jedinců, kteří byli vybráni v předchozím kroku), získáme mnohem větší nárůst hodnoty znaku než při jednorázovém výběru jednoho embrya ze 100. Postupná selekce má pochopitelně tu nevýhodu, že trvá déle. Je-li rozestup mezi dvěma generacemi dvacet nebo třicet let, už jenom po pěti generacích se ocitneme hluboko v 22. století. Dlouho předtím už nejspíš budeme mít k dispozici přímočařejší a efektivnější metody genetického inženýrství (nemluvě o strojové inteligenci).

Existuje však komplementární technologie, která by, vyvinuta v podobě aplikovatelné u člověka, schopnosti preimplantační genetické diagnostiky o mnoho zvětšila. Jde o odvozování životaschopných spermií a vajíček z embryonálních kmenových buněk.[132] Pomocí příslušných technik již bylo u myší vytvořeno plodné potomstvo a u lidí buňky podobné pohlavním buňkám. Než však bude možné výsledky dosažené u zvířat zopakovat u lidí, bude nutné vyřešit obtížné vědecké problémy – a téhož bude zapotřebí, abychom se u linií kmenových buněk vyhnuli epigenetickým abnormalitám. Proto podle jednoho experta bude tato technologie u člověka aplikovatelná možná až „za 10, nebo dokonce 50 let".[133]

Jakmile budou pohlavní buňky odvozené z kmenových buněk dostupné, budou lidské páry disponovat mnohem větší selekční silou. V současné praxi se při oplodnění ve zkumavce obvykle vytváří méně než deset embryí. Se zmíněnou technologií budeme moci z několika darovaných buněk vytvořit prakticky neomezené množství pohlavních buněk, které bude možné spojovat za účelem vytvoření embryí. Následně půjde tato embrya genotypovat nebo sekvenovat a to nejslibnější z nich zvolit pro implantaci. V závislosti na nákladnosti přípravy a vyšetření každého jednotlivého embrya by se selekční síla dostupná párům využívajícím oplodnění ve zkumavce mohla několikanásobně zvýšit.

A co je ještě důležitější, tato technologie by nám dovolila stlačit několikagenerační selekci do období kratšího, než je doba lidského dospívání. Umožnila by nám totiž provádět *iterovanou selekci embryí*, což je procedura, jež by sestávala z následujících kroků:[134]

1. Genotypizace a výběr několika embryí, která vykazují vyšší stupeň požadovaných genetických charakteristik.
2. Extrakce kmenových buněk z těchto embryí a jejich přeměna ve spermie a vajíčka, dozrávající za šest nebo méně měsíců.[135]
3. Spojení nové spermie a vajíčka za účelem vytvoření embrya.
4. Opakování téhož postupu, dokud se nenahromadí velké genetické změny.

Tímto způsobem by bylo za pouhých několik let možné provést selekci u deseti nebo více generací. (Tato procedura by byla časově náročná a nákladná, v principu by ji však stačilo provést jen jednou, místo abychom ji museli opakovat pro každé nové dítě. S využitím buněčných linií vytvořených na konci této procedury bychom mohli vytvářet velké množství vylepšených embryí.)

Jak ukazuje tabulka 5, *průměrná* inteligenční úroveň jedinců počatých tímto způsobem by mohla být velice vysoká: mohla by se rovnat inteligenci nejchytřejšího člověka v dosavadních dějinách lidstva nebo ji o něco přesahovat. Pokud by na světě existovalo velké množství takových jednotlivců (a pokud by zároveň měla odpovídající úroveň světová kultura, vzdělávání, komunikační infrastruktura atd.), mohla by vzniknout kolektivní superinteligence.

Existuje několik faktorů, které dopady této technologie umenší a oddálí. Zaprvé dochází k nevyhnutelnému zdržení, protože vybrané embryo musí nejprve dorůst v dospělého jedince: bude to trvat alespoň 20 let, než vylepšené dítě dosáhne plné produktivity, a ještě déle, než takové děti začnou tvořit významnou část pracovní síly. Navíc i poté, co tato technologie bude plně vyvinuta, bude pravděpodobně zprvu používána málo. Některé země ji možná z morálních nebo náboženských důvodů zcela zakážou.[136] A i tam, kde bude selekce povolena, bude mnoho párů upřednostňovat přirozený způsob početí. Ochota k využívání této metody by se však zvýšila, pokud by s ní byly spojeny jednoznačnější přínosy, například kdyby bylo prakticky jisté, že počaté dítě bude velice talentované a nebude geneticky predisponováno k nemocem. Ve prospěch genetické selekce by hovořily rovněž nižší náklady na zdravotní péči a vyšší

očekávané celoživotní výdělky. Až se tato procedura rozšíří (zvláště mezi společenskou elitou), dojde možná ke kulturnímu obratu směrem k rodičovským normám, podle kterých bude užití selekce známkou osvíceného a zodpovědného páru. Mnozí z těch, jejichž postoj bude zpočátku zdráhavý, třeba nakonec půjdou s proudem, aby jejich děti nebyly znevýhodněny ve srovnání s dětmi přátel a kolegů. Některé země zřejmě budou svým občanům dávat pobídky k využití genetické selekce, aby zmnožily svůj lidský kapitál nebo zvýšily dlouhodobou společenskou stabilitu tím, že mimo vládnoucí vrstvu budou podporovat výběr takových znaků jako povolnost, poslušnost, poddajnost, přizpůsobivost, zbabělost či averze k riziku.

Vliv této technologie na intelektuální schopnosti by závisel také na tom, v jakém rozsahu by dostupná selekční síla byla použita k vylepšení kognitivních znaků (viz tabulku 6). Páry, které se pro nějakou formu selekce rozhodnou, budou muset zvážit, jakým směrem dostupnou selekční sílu zaměří, a inteligenci by do určité míry konkurovaly jiné žádoucí vlastnosti, například zdraví, krása, dobrá povaha či fyzická zdatnost. Iterovaná selekce embryí, která nabízí velkou selekční sílu, by zmenšila potřebnost některých těchto kompromisů, neboť by mohla zároveň probíhat silná selekce ve prospěch většího množství znaků. Tento postup by však vedl k narušení normálního genetického vztahu mezi rodiči a dítětem, což by mohlo v mnoha kulturách negativně ovlivnit poptávku.[137]

S dalším pokrokem genetických technologií možná vznikne možnost syntetizovat genom přesně podle požadavků, takže již nebudou zapotřebí velké zásoby embryí. Syntéza DNA je již nyní rutinní a z velké části automatizovaná biotechnologie, ač zatím není možné syntetizovat celý lidský genom, který by byl použitelný v reprodukčním kontextu (v neposlední řadě kvůli zatím nevyřešeným problémům s epigenetikou).[138] Jakmile však vývoj této technologie pokročí, budeme moci sestrojit embrya tak, aby od svých rodičů obdržela preferovanou kombinaci genů. Zároveň budeme moci do genomu embrya vložit geny, které u žádného z jeho rodičů přítomné nejsou, včetně alel, jež se v populaci vyskytují s malou četností, ale které mohou mít výrazný pozitivní vliv na kognitivní schopnosti.[139]

Tabulka 6 ***Možné dopady genetické selekce při různých scénářích***[140]

Míra přijetí metody / používaná technologie	**„IVF+“ Selekce jednoho embrya ze 2 [4 body IQ]**	**„Agresivní IVF“ Selekce jednoho embrya z 10 [12 bodů IQ]**	**„Vajíčko *in vitro*“ Selekce jednoho embrya ze 100 [19 bodů IQ]**	**„Iterovaná selekce embryí“ [100+ bodů IQ]**
„Okrajová praxe“ ~ 0,25% přijetí	Společenské následky během jedné generace zanedbatelné. Vyvolané spory ve společnosti mají větší význam než přímé dopady.	Společenské následky během jedné generace zanedbatelné. Vyvolané spory ve společnosti mají větší význam než přímé dopady.	Skupina vylepšených jedinců tvoří nepřehlédnutelnou menšinu zaujímající pozice s vysokými nároky na kognitivní schopnosti.	Vylepšení jedinci převládají v řadách elitních vědců, advokátů, lékařů, inženýrů. Intelektuální obroda?
„Výhoda elit“ 10% přijetí	Mírné kognitivní dopady v 1. generaci, společně se selekcí nekognitivních znaků přináší menšině populace znatelné výhody.	Vylepšení jedinci tvoří velkou část harvardských studentů. V druhé generaci převládají v profesích s vysokými nároky na kognitivní schopnosti.	V první generaci vylepšení jedinci převládají v řadách vědců, advokátů, lékařů a inženýrů.	„Post-lidstvo“[141]
„Nová norma“ > 90% přijetí	Poruchy učení jsou u dětí mnohem méně časté. V druhé generaci se množství jedinců s vysokým IQ více než zdvojnásobí.	Významné zlepšení vzdělávacích výsledků a nárůst příjmů. V druhé generaci mnohonásobný nárůst na pravém okraji inteligenčního spektra.	Hrubé hodnoty IQ typické pro význačné vědce jsou v první generaci častější alespoň desetkrát, v druhé generaci několik tisíckrát.	„Post-lidstvo“

Až budeme schopni syntetizovat lidský genom, umožní nám to provádět rovněž genetickou „korekturu“ embryí. (K té bychom se mohli přiblížit také s pomocí iterované selekce embryí.) V současnosti je každý z nás zatížen množstvím genetických mutací: může jít o stovky mutací, které snižují efektivitu různých buněčných procesů.[142] Samy o sobě mají tyto mutace zanedbatelné následky (a proto jsou z genofondu odstraňovány jen pomalu), společně však mohou fungování organismu výrazně zhoršovat.[143] Individuální rozdíly v inteligenci mohou být z velké části dány množstvím a povahou lehce škodlivých alel, které v sobě každý z nás nese. Genetická syntéza by nám umožnila vytvořit takovou verzi genomu daného embrya, která by byla zbavena genetického šumu nahromaděných mutací. Pokud bychom se chtěli vyjadřovat provokativně, mohli bychom říci, že jedinci vytvoření z takových zkorigovaných genomů by byli „lidštější“ než kdokoliv z dnešních lidí, neboť by lidskou přirozenost ztělesňovali méně zkresleným způsobem. Takoví lidé by si nicméně nebyli podobní jako vejce vejci, protože lidé se geneticky neodlišují jenom tím, že v sobě nesou rozdílné škodlivé mutace. Fenotypovým projevem zkorigovaného genomu by však mohla být mimořádná tělesná i duševní konstituce – mohlo by dojít ke zlepšením u takových polygenních znaků, jako jsou inteligence, zdraví, odolnost a vzhled.[144] (Přibližnou analogií by mohly být portréty složené z fotografií několika tváří, na nichž se různé kazy jednotlivců vzájemně vyruší: viz obrázek 6.)

Relevantní mohou být i jiné potenciálně dostupné biotechnologie. Až bude vyvinuto reprodukční klonování lidí, mohli bychom s jeho pomocí replikovat genom mimořádně nadaných jedinců. Dopady této technologie by umenšila skutečnost, že většina rodičů by chtěla být biologicky příbuzná svým dětem. Přesto by tyto dopady mohly být významné, zaprvé protože i poměrně malý nárůst počtu mimořádně nadaných lidí by mohl mít značné následky, a zadruhé z toho důvodu, že některý stát by možná zahájil eugenický program v širším měřítku, třeba tak, že by náhradní matky finančně odměňoval. Postupem času třeba získají na významu také jiné druhy genetického inženýrství, například možná budeme schopni navrhovat nové