Singularita

Vernor Vinge
Department of Mathematical Sciences
San Diego State University

Původní verze tohoto článku byla přednesena na sympóziu VISION-21 sponzorovaném NASA Lewis Research Center a Ohio Aerospace Institute konaném 30. – 31. 3. 1993. Poněkud upravená verze se objevila v zimním vydání *Whole Earth Review 1993.*

Abstrakt

Do 30 let budeme mít technické prostředky, abychom vytvořili nadlidskou inteligenci. Krátce poté skončí lidská éra. Je možno tomuto vývoji zabránit? Pokud ne, můžeme řídit běh událostí tak, abychom dokázali přežít? Úvaha se zabývá těmito otázkami a přináší možné odpovědi (i některá další nebezpečí).

Co je to Singularita?

Ústředním rysem tohoto století je zrychlování technického rozvoje. V tomto článku dovozuji, že jsme na pokraji změny srovnatelné s objevením člověka na Zemi. Vlastní příčinou této změny je to, že v brzké době vytvoříme bytosti [V originále se mluví o "entitách". Protože v češtině je slovo "entita" mnohem méně běžné a pociťuji je jako termín se silně filozofickým podtextem, překládám raději jako "bytost" – pozn. překl.] s inteligencí vyšší, než je lidská. Tohoto průlomu může věda dosáhnout více způsoby (a to je další důvod, proč si můžeme být jisti, že k tomu dojde).

- Mohou být vyvinuty počítače, které budou mít vědomí a nadlidskou inteligenci. (Doposud je sporné, zda můžeme vytvořit strojový ekvivalent člověka. Pokud však odpověď zní "ano", lze sotva pochybovat o tom, že krátce poté mohou být vytvořeny bytosti inteligentnější.)
- Rozsáhlé počítačové sítě (a jejich uživatelé) se mohou "probudit" coby bytosti s nadlidskou inteligencí.

• Propojení (interfejs) člověka s počítačem se může stát tak těsným, že uživatele bude možno považovat za tvory s nadlidskou inteligencí.
• Biologie může poskytnout prostředky pro zvýšení přirozené lidské inteligence.

První tři možnosti zásadně závisí na zdokonalení hardwaru počítačů. Rozvoj v této oblasti sleduje v posledních desetiletích stabilní křivku [17]. Na základě tohoto trendu věřím, že k vytvoření inteligence vyšší, než je lidská, dojde v průběhu následujících třiceti let. (Charles Platt [20] poukázal na to, že nadšenci pro umělou inteligenci – AI – vyslovují taková tvrzení už 30 let. Abych časové zařazení upřesnil, budu konkrétnější: Budu překvapen, stane-li se to před rokem 2005 nebo po roce 2030.)

Jaké bude mít tato událost důsledky? Až se nadlidská inteligence stane hybnou silou rozvoje, bude tento rozvoj mnohem rychlejší. Není vlastně žádný důvod, proč by tento rozvoj nevedl v ještě kratší době k vytvoření ještě inteligentnějších bytostí. Nejlepší analogie, která mě napadá, je evoluce v minulosti: Zvířata se mohou přizpůsobit problémům a dělat vynálezy, ale často ne rychleji, než pracuje přirozený výběr – v případě přirozeného výběru je svět svým vlastním simulátorem. My lidé jsme schopni obsáhnout svět a ve své hlavě analyzovat "co se stane, když..."; mnoho problémů jsme schopni vyřešit tisíckrát rychleji než přirozený vývoj. Vytvoříme-li nyní prostředky, které budou takové simulace provádět mnohem rychleji, dostáváme se do režimu, který se od naší lidské minulosti bude lišit stejně, jako my lidé od zvířat.

Z lidského hlediska bude tato změna znamenat odhození všech dosavadních pravidel, možná v jediném okamžiku, exponenciální lavinu vymykající se zcela jakékoli kontrole. K čemu jsme mysleli, že může dojít "za milion let" (pokud vůbec), to se pravděpodobně stane v příštím století. (V [5] Greg Bear kreslí obrázek závažných změn, které proběhnou během několika hodin.)

Myslím, že takovou událost můžeme směle nazvat *Singularitou* (pro účely tohoto článku ***Singularitou*** s velkým "S"). Je to okamžik, kdy musíme odhodit své staré modely a kdy začne vládnout nová realita. Čím více se budeme blížit tomuto okamžiku, tím více bude tato realita ovlivňovat naše záležitosti, dokud to nebude zcela zřejmé. Přesto, až se to nakonec stane, může to být velkým překvapením a ještě větší neznámou. V padesátých letech tuto budoucnost vidělo jen velmi málo lidí – Stan Ulam [28] parafrázoval Johna von Neumanna: Jeden rozhovor se soustředil na stále se zrychlující technický pokrok a změny v životním stylu lidí, což nasvědčuje tomu, že se blíží zásadní singularita v historii lidstva, po níž lidské záležitosti, jak je známe, nebudou moci pokračovat.

Von Neumann dokoce použil termín *singularita*, i když měl zřejmě na mysli normální vývoj a ne vytvoření nadlidského intelektu. (Podle mne je nadlidskost podstatou Singularity. Bez ní bychom došli k nadbytku technického bohatství, které bychom nikdy nedokázali využít - viz [25].)

V 60. letech byly pochopeny některé důsledky nadlidské inteligence. I. J. Good napsal [11]:

> *„Definujme ultrainteligentní stroj jako stroj, který daleko předčí intelektuální činnosti sebechytřejšího člověka. Protože vývoj strojů je jednou z těchto intelektuálních činností, ultrainteligentní stroj může vyvíjet ještě lepší stroje; dojde nepochybně k "inteligenční explozi" a lidská inteligence zůstane daleko vzadu. První ultrainteligentní stroj bude proto posledním vynálezem, který člověk udělá, pokud tento stroj bude dostatečně hodný na to, aby nám řekl, jak ho můžeme udržet pod kontrolou. (...) Je pravdě-podobné na více než 50 %, že ještě ve 20. století bude ultrainteligentní stroj postaven a že to bude poslední vynález, který člověk bude muset udělat.“*

Good postřehl podstatu valící se laviny, ale nevěnoval se jejím nejvíce znepokojivým důsledkům. Žádný inteligentní stroj toho druhu, jaký popisuje, by se nestal "nástrojem" lidstva - tak jako lidé nejsou nástroji králíků, drozdů nebo šimpanzů.

V průběhu 60., 70. a 80. let se šířilo rozpoznání pohromy [29], [1], [31], [5]. Asi jako první vycítili její konkrétní dopad autoři science fiction. Konec konců právě autoři "tvrdé" science fiction psali příběhy o všem, co pro nás technika může udělat. Stále více tito spisovatelé cítili neprostupnou zeď táhnoucí se budoucností. Kdysi umísťovali takové fantazie do milionů let vzdálené budoucnosti [24]. Nyní vidí, že jejich odvážné extrapolace vedou do budoucnosti neznámo jak ... blízké. Kdysi mohla galaktická impéria vypadat jako záležitost posthumánní éry. Nyní do ní bohužel patří i říše meziplanetární.

A co léta devadesátá a nultá a desátá, jak se blížíme k hranici? Jak se projeví přibližování Singularity na lidském pohledu na svět? Ještě chvíli se bude kritikům strojového rozumu popřávat sluchu. Ostatně, dokud nebudeme mít hardware s výkonností lidského mozku, je asi pošetilé myslet si, že bychom mohli vytvořit inteligenci ekvivalentní lidské (nebo vyšší). (Je zde jistá za vlasy přitažená možnost, že bychom vytvořili lidský ekvivalent z méně výkonného hardwaru, pokud bychom chtěli obětovat rychlost, kdybychom byli ochotni se spokojit s umělou bytostí, která by byla doslova opožděná [30]. Je však mnohem pravděpodobnější, že vytvoření takového softwaru by bylo choulostivý problém s mnoha slepými uličkami a mnoha experimenty. V takovém případě by se stroje s vědomím objevily za tak dlouho, že by se do té doby objevil hardware podstatně výkonnější než ten lidský.)

V průběhu času bychom však měli vidět další symptomy. Dilema pociťované autory science fiction začnou vnímat i lidé pracující v jiných tvůrčích oblastech. (Slyšel jsem, jak si hloubaví autoři comicsů

stěžují, jak je těžké vymýšlet si působivé efekty, když všechno, co lze nakreslit, je možné vytvořit dnes už běžnou technikou.) Staneme se svědky toho, jak automatizace bude nahrazovat lidi na složitějších a složitějších pracovních místech. Už dnes máme nástroje (symbolické matematické výpočty, CAD-CAM), které nás zbavují většiny jednoduché otročiny. Jinak řečeno, skutečně produktivní práce je doménou stále menší a elitnější části lidstva. V příchodu Singularity vidíme, že předvídaná *skutečná* technická nezaměstnanost se s konečnou platností stane skutkem.

Další příznak blížící se singularity: samotné myšlenky se budou šířit ještě rychleji a i ty nejradikálnější se rychle stanou samozřejmými pravdami. Když jsem uprostřed 60. let začal psát science fiction, připadalo mi velmi snadné najít nápady, které teprve po desetiletích pronikly do kulturního povědomí; dnes se realizují za nějakých osmnáct měsíců. (Samozřejmě může být na vině ztráta mé představivosti spojená s mým stárnutím, ale tento efekt pozoruji i na ostatních.) Jak nabíráme rychlost přes kritickou hodnotu, Singularita se blíží jako rázová vlna ve stlačitelné kapalině.

A co příchod Singularity samotné? Co se dá říci o tom, jak bude doopravdy vypadat? Protože se zakládá na lavinovitém růstu inteligence, proběhne pravděpodobně rychleji než jakákoli dosavadní technická revoluce. Bleskový účinek bude asi neočekávaný - možná i pro vědce, kteří se na něm budou podílet. ("Vždyť všechny naše předchozí modely byly netečné! Jenom jsme vyšolíchali pár parametrů...") Budou-li dostatečně rozvinuté sítě (budou zahrnovat i všudypřítomné počítače vestavěné do jiných zařízení), může se zdát, jako by se naše artefakty jako celek náhle probudily.

A co se stane za jeden dva měsíce (nebo za jeden dva dny)? Mohu ukázat pouze na jednu analogii: vznik lidstva. Budeme v posthumánní éře. A při všem mém nezkrotném technickém optimismu si někdy myslím, že bych se cítil lépe, kdybych uvažoval o těchto událostech jako o něčem vzdáleném tisíc let – a ne dvacet.

Je možné se Singularity vyvarovat?

No, možná k tomu vůbec nedojde: někdy se pokouším představit si symptomy, které bychom mohli očekávat, kdyby se Singularita neměla vyvinout. Široce jsou přijímány argumenty Penrosa [19] a Searla [22] proti praktické možnosti vytvoření strojového rozumu. V srpnu 1992 uspořádala společnost Thinking Machines Corporation seminář (workshop) na téma "Jak sestrojíme stroj, který myslí" [27]. Jak lze usoudit z názvu, účastníci nebyli příliš nakloněni argumentům proti strojové inteligenci. Panoval prakticky souhlas s názorem, že myšlení může být založeno na nebiologickém základě a že zásadní význam pro existenci myšlení mají algoritmy. Mnoho se však debatovalo o tom, jak výkonný je hardware lidského mozku. Většina účastníků souhlasila s Moravcovým [17] odhadem, podle něhož

nás od vyrovnání hardwarového výkonu dělí deset až čtyřicet let. Přesto zde byla zastoupena menšina, která poukazovala na [7], [21] a soudila, že výpočetní mohutnost jednotlivých neuronů může být mnohem větší, než se běžně věří. Je-li to pravda, mohlo by současnému počítačovému hardwaru do zařízení, které nosíme v hlavě, chybět *deset* řádů. Pokud je tomu tak (nebo je-li Penrosova či Searlova kritika oprávněná), Singularitu možná nikdy neuvidíme. Místo toho bychom v prvním desetiletí po roce 2000 mohli zjistit, že se výkonová křivka hardwaru začíná narovnávat – v důsledku naší neschopnosti automatizovat návrhářskou práci potřebnou k vývoji dalších hardwarových zlepšení. Skončili bychom u jistého *velmi* výkonného hardwaru, ale bez možnosti rozvíjej jej dále. Komerční digitální zpracování signálu by mohlo budit velký respekt a mohlo by se zdát, že i u digitálních operací jde o zpracování analogové, ale nikdy by se nic "neprobudilo" a nedošlo by k lavinovitému rozvoji inteligence, v němž tkví podstata Singularity. Pravděpodobně by tento stav byl považován za zlatý věk a byl by to současně konec pokroku. Velice se to podobá budoucnosti předpovězené Guntherem Stentem. V [25, s. 137] Stent výslovně uvádí vyvinutí nadlidské inteligence jako dostatečnou podmínku pro nenaplnění jeho předpovědí.

Pokud však k technické singularitě může dojít, stane se tak. Dokonce, i kdyby všechny vlády světa chápaly "hrozbu" a bály se jí jako čert kříže, vývoj by dále směřoval k tomuto cíli. V beletrii se objevily příběhy o zákonech, které zakazovaly konstrukci "stroje k obrazu lidské mysli" [13]. Konkurenční převaha - ekonomická, vojenská, dokonce i umělecká – každého pokroku v automatizaci je vskutku tak přesvědčivá, že přijímáni zákonů či uvalení cla, které by takové věci znemožňovaly, pouze zajistí, že cíle dosáhne dříve někdo jiný.

Eric Drexler [8] působivě pronikl do podstaty toho, jak daleko může jít technický rozvoj. Souhlasí s tím, že nadlidské inteligence budou k dispozici v blízké budoucnosti – a že takové bytosti představují ohrožení lidského *status quo*. Drexler však tvrdí, že taková nadlidská zařízení můžeme udržet na uzdě, takže bychom jejich výsledky mohli bezpečně zkoumat a používat. Je to ultrainteligentní stroj I. J. Gooda s jistou dávkou obezřetnosti. Já tvrdím, že takové zkrocení je v praxi zásadně neproveditelné. V případě fyzického uvěznění: představte si, že jste zamčeni ve svém domě s pouze omezeným datovým spojením ven ke svým pánům. Pokud tito páni myslí řekněme milionkrát pomaleji než vy, sotva můžeme pochybovat, že během několika let (vašeho času) byste mohli přijít s "užitečnou radou", která vás jen tak mimochodem osvobodí. (Říkám tomu "rychle myslící" forma superinteligence, "slabá nadlidskost". Taková "slabě nadlidská" bytost by se pravděpodobně "prokopala" na svobodu během několika týdnů vnějšího času. "Silná nadlidskost" by byla víc než jen ekvivalent lidské mysli roztočený na vysoké obrátky. Je těžké říci přesně, jak by "silná nadliskost" vypadala, ale rozdíly by asi byly zásadní. Představte si velmi rychlé psí myšlení. Pomohly by tisíce let psího života lidskému chápání?

(No, kdyby se psí myšlení šikovně "předrátovalo" a *potom* rozjelo vysokou rychlostí, to by mohlo být něco jiného...) Zdá se, že mnoho spekulací o superinteligenci je založeno na "slabě nadlidském" modelu. Věřím, že postsingulární svět odhadneme nejlépe tím, že budeme přemýšlet nad vlastnostmi silné nadlidskosti. K tomu se v další části ještě vrátím.)

Jinou možností je vestavět do mysli vytvářené nadlidské bytosti *pravidla* (například Asimovovy zákony robotiky [3]). Myslím si, že každá pravidla tak přísná, aby byla účinná, vytvoří současně zařízení, jehož schopnosti budou jednoznačně pokulhávat za nespoutanými verzemi (a lidská soutěživost tak bude dávat přednost nebezpečnějším modelům). Přesto je Asimovův sen úžasný: představte si ochotného otroka, který má po všech stránkách tisícinásobek vašich schopností. Představte si stvoření, které může splnit každé vaše bezpečné přání (ať tento termín znamená cokoli) a stále mu zbývá 99,9 % pro jiné činnosti. To by byl nový svět, který bychom nikdy opravdu nechápali, ale byl by plný laskavých bohů (i když jedno z *mých* přání by bylo stát se jedním z nich).

Pokud Singularitě nemůžeme zabránit ani ji spoutat, jak zlá by mohla být posthumánní éra? No hodně zlá. Jednou z možností je vyhynutí lidské rasy. (Vezmeme-li v úvahu, co všechno taková technika může dokázat, vlády by mohly přijít k závěru, že občany už nepotřebují!) Přesto by fyzické vyhubení nemuselo být tou nejděsivější možností. Použijme opět analogii: Pomysleme na naše chování ke zvířatům. Některé způsoby krutého fyzického týrání jsou nepřijatelné, ale...

V posthumánním světě by stále byla spousta míst, kde je žádoucí automatizace vyžadující schopnosti ekvivalentní lidským: systémy začleněné do autonomních zařízení, vědomím vybavené kontrolní subsystémy pro výkon nižších činností ve větších cítících organizmech.) Některé z těchto lidských ekvivalentů by mohly být využívány pro pouhé digitální zpracování signálů. Byly by podobnější velrybám než lidem. Jiné by mohly vypadat velice lidsky, ale byly by jednostranné - natolik, že dnes by je to přivedlo na psychiatrickou kliniku. Třebaže žádní z těchto tvorů by nemuseli být lidmi z masa a kostí, mohli by v novém prostředí být čímsi, co je nejbližší k tomu, co dnes nazýváme lidmi. (I. J. Good k tomu měl co říci, ačkoli dnes už je jeho rada diskutabilní. Good [12] navrhl "meta-zlaté pravidlo", které lze parafrázovat slovy: "Chovej se k nižším tvorům tak, jak bys chtěl, aby se vyšší tvorové chovali k tobě." Je to skvělá, byť paradoxní idea - a většina mých přátel v ni nevěří - protože očekávaná výhra ve smyslu teorie her se obtížně vyjadřuje. Kdybychom to však dokázali, mohla by nám v jistém smyslu říci něco o pravděpodobnosti takové dobrosrdečnosti v tomto vesmíru.)

V předchozím jsem argumentoval, že Singularitě nemůžeme zabránit, že se blíží jako nevyhnutelný důsledek lidské přirozené soutěživosti a možností technice vlastní. A přesto - my jsme to uvedli do pohybu. I ta největší lavina se spouští maličkostmi. Máme svobodu v tom, jaké počáteční podmínky

zavedeme, aby se věci udály co nejméně nepřátelsky. Samozřejmě (jako při spouštění laviny) nemusí být jasné, kterým směrem máme věci "pošťouchnout".

Další cesty k singularitě: posilování inteligence

Mluví-li lidé o vytvoření nadlidsky inteligentních bytostí, představují si obvykle projekty umělé inteligence (AI). Jak jsem však poznamenal na začátku této stati, k nadlidskosti vedou i jiné cesty. Počítačové sítě a interfejs člověk-počítač vypadají pozemštěji než AI, a přesto mohou vést k Singularitě. Toto odlišné pojetí budu označovat jako posilování inteligence (intelligence amplification – IA). IA je něco, co postupuje velmi přirozeně, většinou ti, kdo ho vyvíjejí, ani nepostřehnou, co se děje. Avšak pokaždé, kdy se zlepší naše schopnost přístupu k informacím a předáme ji ostatním, udělali jsme v jistém smyslu krůček nad přirozenou inteligenci. Už dnes tým složený z lidí s vědeckým titulem a s dobrou pracovní stanicí (i když nebude zapojena do sítě!) by asi mohl dosáhnout špičkových výsledků ve kterémkoli psaném testu inteligence.

A je velmi pravděpodobné, že IA je mnohem snazší cesta k dosažení nadlidskosti než čistá AI. V lidských bytostech byly největší vývojové problémy už asi vyřešeny. Stavět na tom, co je v nás, by mělo být snazší než nejprve zjistit, co jsme zač, a potom konstruovat stroje na stejném principu. A je zde aspoň hypotetický precedens. Cairns-Smith [6] spekuloval, že biologický život možná začal jako přídavek k ještě primitivnějšímu životu založenému na růstu krystalů. Lynn Margulis (v [15] i jinde) předložil pádné argumenty ve prospěch názoru, že vzájemný vztah je v evoluci silným hybným činitelem.

Poznamenávám, že nenavrhuji, abychom ignorovali výzkum AI nebo na něj věnovali méně prostředků. To, co se povede v AI bude asi použitelné i v IA a naopak. Navrhuji pouze uvědomit si, že ve výzkumu sítí a interfejsů je něco stejně závažného (a potenciálně nebezpečného) jako v AI. S tímto pochopením před sebou můžeme vidět projekty, které nejsou jen přímo aplikovatelné jako konvenční interfejsy a návrh sítí, ale které nás posunou blíže k Singularitě po cestě IA.

Zde jsou některé možné projekty, které mají z hlediska IA zvláštní význam:

- Automatizace týmu člověk-počítač: Vezměme si problémy, které se normálně považují za vhodné pro čistě strojové řešení (jako je problém maximalizace funkcí) a navrhněme programy a interfejs, které využijí lidskou intuici a počítačový hardware. Vezmeme-li v úvahu všechnu bizarnost vícerozměrných problémů tohoto typu (a elegantní algoritmy, které byly pro jejich řešení vymyšleny), bylo by možno navrhnout velmi zajímavé zobrazovací a ovládací prvky pro použití lidským členem týmu.

- Vývoj symbiózy člověk-počítač v umění: Zkombinujme grafické zobrazovací možnosti moderních počítačů a estetické cítění člověka. Samozřejmě již bylo věnováno ohromné úsilí návrhu počítačových nástrojů pro umělce zaměřené na usnadnění jejich práce. Navrhuji, abychom se výslovně zaměřili na větší splynutí úloh člověka a počítače, abychom explicitně posílili kooperatovní přístup. Obdivuhodnou práci v tomto směru udělal Karl Sims [23].
- Dovolit na šachových turnajích kombinované týmy člověk-počítač. Už máme programy, které dokáží hrát lépe než většina lidí. [Pozn. překl.: Od roku 1997 můžeme vynechávat slovo "většina" – program Deep Blue běžící na počítači doplněném speciálním šachovým hardwarem porazil v zápase na 6 partií úřadujícího mistra světa. Brzy poté byl počítač ve zmíněné speciální úpravě likvidován.] Kolik úsilí však bylo věnováno tomu, jak by jejich sílu mohl využít člověk, abychom získali něco ještě lepšího. Kdyby účast takových týmů byla připuštěna alespoň na některých turnajích, mohlo by to na výzkum IA mít pozitivní efekt, stejně jako připuštění účasti samotných počítačů mělo v oblasti AI.
- Vyvinout interfejs, který by umožňoval přístup k počítači a k síti, aniž by člověk byl uvázán na jediné místo vsedě před počítačem. (Tento aspekt IA představuje tak velké ekonomické výhody, že je mu už věnováno mnoho úsilí.)
- Vytvořit symetričtější systémy pro podporu rozhodování. Výzkum a tvorba rozhodovacích systémů je v posledních letech populární oblastí. Je to forma IA, ale možná je příliš soustředěna na systémy "věštecké". Stejně důležité jako to, aby systém dával uživateli informace, musí být i otázka, jak má uživatel dávat systému vodítka.
- Využívat lokální sítě tak, aby lidské týmy skutečně pracovaly (tj. aby byly efektivnější než jejich samostatní členové). Obecně to spadá do oblasti "groupware", což už je velmi populární komerční činnost. Přístup k ní by se však měl zaměřit na to, aby činnost skupiny byla chápána jako sladěný organismus. V jednom smyslu lze tento návrh brát s cílem vytvořit "pravidla uspořádání" takových sladěných organismů. Například hlavní zaměření může být udržováno snáze než pomocí klasických pracovních porad. Kvalifikace jednotlivých členů týmu může být oddělena od osobních záležitostí, aby se přínos jednotlivých členů soustředil na týmový projekt. A samozřejmě sdílené databáze se využívají mnohem lépe než výsledky činnosti komise. (Poznamenávám, že tento návrh je zaměřen na týmovou práci a ne na politické schůze. V politice by výše popsaná automatizace pouze zvýšila moc těch, kdo určují pravidla!)
- Využívat celosvětově rozšířený Internet jako nástroj pro souhru člověka a stroje. Ze všech bodů tohoto seznamu rozvoj tohoto bodu postupuje nejrychleji a může nás dovést k Singularitě dříve než cokoli jiného. Síla a vliv i dnešního Internetu [tj. Internetu z r. 1993 – pozn. překl.] se výrazně podceňuje. Například si myslím, že naše současné počítačové systémy by se zhroutily pod tíhou vlastní složitosti nebýt výhody, kterou správcům systému a lidem z technické podpory dává "skupinové vědomí" Usenetu! Sama anarchie panující v celosvětovém vývoji sítí je důkazem jejího potenciálu. Jak

se rozšiřuje konektivita, přenosová rychlost, rozsah archivů a rychlost počítačů, vidíme něco jako Margulisovu [15] vizi biosféry jakožto vyššího stupně datového procesoru, milionkrát rychlejšího a složeného z milionů lidských inteligentních komponent (nás).

Výše uvedené příklady ilustrují výzkum, který může probíhat v rámci současných kateder výpočetní techniky (informatiky). Jsou i jiná paradigmata. Například mnohým pracím spojeným s umělou inteligencí a neuronovými sítěmi by prospělo těsnější sepětí s biologickým životem. Místo pokusů prostě modelovat biologický život na počítačích a pochopit ho by výzkum mohl být zaměřen na vytvoření složitých systémů spoléhajících na biologický život jako vodítko nebo pro využití vlastností, které dosud neznáme dostatečně na to, abychom je implementovali v hardwaru. Dlouhotrvajícím snem science fiction je přímé napojení mozku na počítač [2], [29]. Skutečně existují konkrétní práce, které je možno v této oblasti provádět (a které se již provádějí):

- Protézy končetin jsou oblastí s přímou komerční použitelností. Transducery nerv-křemík jsou vyrobitelné [14]. Jde o vzrušující nepříliš vzdálený krok k přímé komunikaci.
- Přímé propojení do mozku vypadá dosažitelně, pokud bude nízká přenosová rychlost. Při lidské flexibilitě procesu učení by nemuselo být nutné vybrat přesný cíl. I rychlost 1000 bit/s by bylo velkou pomocí pro oběti mozkové mrtvice, které by jinak byly odkázány na interfejs založený na menu.
- Připojení na zrakový nerv by umožnilo dosáhnout přenosové rychlosti kolem 1 Mbit/s. K tomu bychom však potřebovali znát detailně, jak funguje zpracování zrakových vjemů, a museli bychom umístit obrovskou síť elektrod s mimořádnou přesností. Pokud by naše rychlé připojení mělo fungovat *navíc* k drahám, které už v mozku jsou, problém se stává mnohem nezvládnutelnější. Prosté zasunutí sítě propojovacích prvků do mozku by určitě nefungovalo. Předpokládejme však, že by taková síť s rychlým přenosem byla v mozku přítomna už v době, kdy se struktura mozku utváří během vývoje embrya. Z toho vyplývají:
- Pokusy na zvířecích embryích. V prvních letech takového výzkumu bych nečekal žádné úspěchy na poli IA, ale poskytnout vyvíjejícím se mozkům přístup ke komplexně simulovaným neuronovým strukturám by mohlo být velmi zajímavé pro lidi, kteří studují, jak se vyvíjí mozek embrya. V dlouhodobém výhledu by takové experimenty mohly vést k vytvoření živočichů s přidanými senzorickými drahami a zajímavými intelektuálními schopnostmi.

Původně jsem doufal, že tato diskuse o IA vyústí v nějaký zřetelně bezpečnější přístup k Singularitě. (Konec konců IA připouští, abychom se podíleli na jakési transcendentnosti.) Bohužel, podívám-li se zpět na zmíněné návrhy pro IA, skoro vše, čím jsem si jist, je to, že by se měly vzít v úvahu, že by nám mohly poskytnout další možnosti. Pokud se však týká bezpečnosti – některé z těchto návrhů jsou poněkud děsivé. Jeden z mých neformálních recenzentů poukázal na to, že IA pro jednotlivé lidi

vytváří dost zlověstnou elitu. My lidé jsme zatíženi miliony let evoluce, díky níž se díváme na soupeření druhů v mrtvolném osvětlení. Mnohé takové obavy nejsou nutné v dnešním světě, kde ten, kdo prohrává, přebírá finty vítěze a je kooptován do vítězné firmy. Bytost stvořená *de novo* by mohla být mnohem dobrotivější než ta, v jejíchž kořenech byly tesáky a drápy. A i na rovnostářskou vizi Internetu, který se probudí spolu s celým lidstvem, se lze dívat jako na noční můru [26].

Problém není prostě v tom, že Singularita představuje ústup lidstva ze středu scény, ale v tom, že protiřečí našim hluboce zakořeněným představám o existenci. Myslím, že bližší pohled na pojem silné nadlidskosti může vysvětlit, proč tomu tak je.

Silná nadlidskost a to nejlepší, co můžeme chtít

Předpokládejme, že můžeme Singularitu přizpůsobit. Předpokládejme, že můžeme dosáhnout splnění našich nejpřemrštěnějších nadějí. Co si potom můžeme přát – aby se našimi nástupci stali zase lidé, aby se každá nespravedlnost zmírnila naší znalostí vlastních kořenů. Pro ty, kdo zůstanou nezměněni, by cílem bylo vlídné zacházení (možná i s tím, že zaostalí obyčejní lidé budou mít dojem, že jsou pány božských otroků). Mohl by to být zlatý věk, v němž by existoval pokrok (přeskočení Stentovy bariéry). Bylo by možno dosáhnout nesmrtelnosti (nebo aspoň délky života limitované existencí vesmíru [10], [4]).

V tomto nejradostnějším a nejlaskavějším světě by nastaly hrozivé problémy filozofického rázu. Mysl, jejíž kapacita zůstává stálá, nemůže žít věčně; po několika tisících let by se podobala více smyčce magnetofonové pásky než osobnosti. (Z toho, co jsem na toto téma četl, mi nejvíce běhal mráz po zádech v [18].) Aby se dalo žít neomezeně dlouho, musí se samotná mysl rozvíjet, a když se stane dostatečně velkou, jaký pocit sounáležitosti bude mít s duší, kterou původně bývala? Nová bytost by samozřejmě byla vším, čím byl originál, ale současně něčím tolik větším. A tak i pro individuum musí stále platit Cairnsova-Smithova nebo Margulisova představa nového života inkrementálně vyrůstajícího ze starého.

Tento "problém" s nesmrtelností se vynořuje mnoha dalšími bezprostředními způsoby. Pojem ega a sebeuvědomění byl základem střízlivého racionalismu posledních několika století. Ještě dnes na pojem vědomí útočí pracovníci z oblasti umělé inteligence ("sebeuvědomění a jiné bludy"). Posilování inteligence podkopává naše pojetí ega z jiné strany. Postsingulární svět by vyžadoval sítě s extrémně velkou rychlostí (šířkou pásma). Ústředním rysem silně nadlidských bytostí by pravděpodobně byla jejich schopnost komunikace s proměnnou šířkou pásma zahrnující i rychlosti podstatně větší než u řeči nebo psaného slova. Co se stane, když kousky ega bude možno zkopírovat a propojit a úroveň sebeuvědo-

mění se bude moci rozšiřovat nebo zužovat podle odpovídajících problémů? To jsou základní rysy silné nadlidskosti a Singularity. Přemýšlíme-li o nich, začneme vyciťovat, jak zásadně divná a odlišná bude posthumánní éra – *byť by byla vytvořena sebechytřeji a sebeneškodněji*.

Na jedné straně předložená vize splňuje mnoho našich nejšťastnějších snů: nekonečný čas, kde se můžeme navzájem opravdu poznat a pochopit nejhlubší tajemství. Na straně druhé, se to hodně podobá scénáři nejhoršího případu, který jsem předestřel v dřívější části tohoto článku.

Jaké hledisko platí? Ve skutečnosti si myslím, že nová éra je prostě příliš odlišná, než aby se mohla zařadit do klasického schématu dobra a zla. Tyto pojmy předpokládají izolované neměnné mysli komunikující navzájem málo a pomalu. Postsingulární svět však *zapadá* do širší tradice změn a spolupráce, která začala před dávnou dobou (možná dokonce před vznikem biologického života). Myslím, že *existují* etické pojmy, které by se v takové éře daly aplikovat. Výzkum v oblasti IA a rychlých komunikací by měl toto porozumění zlepšit. Záblesky vidím už dnes [32]. Je to Goodovo meta-zlaté pravidlo; možná existují pravidla, která odliší "sebe" od "druhých" na základě rychlosti spojení. A zatímco myšlení a "vlastní já" budou mnohem vágnější pojmy, mnohé z toho, čeho si ceníme (vědomosti, paměť, myšlenky), se nemusí nikdy vytratit. Myslím, že Freeman Dyson to vystihl, když říká [9]: *"Bůh je to, čím se stane mysl, až se vymkne možnostem našeho chápání."*

Chtěl bych poděkovat Johnu Carrollovi ze San Diego State University a Howardu Davidsonovi ze Sun Microsystems za to, že se mnou prodiskutovali předběžný text tohoto článku.

Anotované prameny (a žádost o případnou bibliographickou pomoc)

[1] Alfvén, Hannes, píšící pod jménem Olof Johanneson, *The End of Man?* (*Konec člověka?*), Award Books, 1969; předtím publikováno pod názvem "The Tale of the Big Computer", Coward-McCann, přeloženo z knihy s copyrightem 1966 Albert Bonniers Forlag AB a copyrightem anglického překladu 1966 Victor Gollanz, Ltd.

[2] Anderson, Poul, "Kings Who Die" (Králové, kteří umírají), If , březen 1962, str. 8-36. Přetištěno v Seven Conquests, Poul Anderson, MacMillan Co., 1969.

[3] Asimov, Isaac, "Runaround", Astounding Science Fiction, březen 1942, str. 94. Přetištěno v Robot Visions, Isaac Asimov, ROC, 1990. V této knize Asimov popisuje vývoj svých robotických příběhů.

[4] Barrow, John D. a Tipler, Frank J., The Anthropic Cosmological Principle (Antropický kosmologický princip), Oxford University Press, 1986.

[5] Bear, Greg, "Blood Music" (Hudba krve), Analog Science Fiction - Science Fact , červen 1983. Rozšířeno na stejnojmenný román, Morrow, 1985.

[6] Cairns-Smith, A. G., Seven Clues to the Origin of Life (Sedm klíčů ke vzniku života), Cambridge University Press, 1985.

[7] Conrad, Michael et al., "Towards an Artificial Brain" (K umělému mozku), BioSystems, sv. 23, str. 175-218, 1989.

[8] Drexler, K. Eric, Engines of Creation (Tvořivé stroje), Anchor Press/Doubleday, 1986.

[9] Dyson, Freeman, Infinite in All Directions (Nekonečné ve všech směrech), Harper && Row, 1988.

[10] Dyson, Freeman, "Physics and Biology in an Open Universe" (Fyzika a biologie v otevřeném vesmíru), Review of Modern Physics, sv. 51, str. 447-460, 1979.

[11] Good, I. J., "Speculations Concerning the First Ultraintelligent Machine" (Spekulace o prvním ultrainteligentním stroji), v Advances in Computers, sv. 6, ed. Franz L. Alt a Morris Rubinoff, str. 31-88, 1965, Academic Press.

[12] Good, I. J., [**Pomozte!** Nemohu najít původní pramen Goodova meta-zlatého pravidla, i když si jasně vzpomínám, že jsem o něm slyšel někdy v 60. letech. Pomocí Internetu jsem našel odkazy na mnoho příbuzných článků. G. Harry Stine a Andrew Haley psali, jak by se metazákon mohl vztahovat na mimozemšťany: G. Harry Stine, "How to Get along with Extraterrestrials ... or Your Neighbor", Analog Science Fact- Science Fiction, únor 1980, str. 39-47.]

[13] Herbert, Frank, Dune (Duna), Berkley Books, 1985. Román však vycházel na pokračování v 60. letech v Analog Science Fiction-Science Fact. [Pozn. překl.: I knižně vyšel poprvé dříve než v r. 1985.]

[14] Kovacs, G. T. A. et al., "Regeneration Microelectrode Array for Peripheral Nerve Recording and Stimulation" (Regenerační systém mikroelektrod pro záznam činnosti a stimulaci periferních nervů), IEEE Transactions on Biomedical Engineering, sv. 39, č. 9, str. 893-902.

[15] Margulis, Lynn a Dorion Sagan, Microcosmos, Four Billion Years of Evolution from Our Microbial Ancestors (Mikrokosmos, čtyři miliardy let vývoje od našich mikrobiálních předků), Summit Books, 1986.

[16] Minsky, Marvin, Society of Mind (Společnost ducha), Simon and Schuster, 1985.

[17] Moravec, Hans, Mind Children (Děti ducha), Harvard University Press, 1988.

[18] Niven, Larry, "The Ethics of Madness" (Etika šílenství), If, duben 1967, str. 82-108. Přetištěno v Neutron Star, Larry Niven, Ballantine Books, 1968.

[19] Penrose, Roger, The Emperor's New Mind (Císařova nová mysl), Oxford University Press, 1989.

[20] Platt, Charles, soukromé sdělení.

[21] Rasmussen, S. et al., "Computational Connectionism within Neurons: a Model of Cytoskeletal Automata Subserving Neural Networks", v Emergent Computation, ed. Stephanie Forrest, str. 428-449, MIT Press, 1991.

[22] Searle, John R., "Minds, Brains, and Programs" (Myšlení, mozky a programy), v The Behavioral and Brain Sciences, sv. 3, Cambridge University Press, 1980. Článek je přetištěn v The Mind's I, editoři Douglas R. Hofstadter a Daniel C. Dennett, Basic Books, 1981 (mnou použitý pramen). Tento reprint obsahuje skvělou kritiku Searlova článku.

[23] Sims, Karl, "Interactive Evolution of Dynamical Systems" , Thinking Machines Corporation, Technical Report Series (publikováno v Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life, Paris, MIT Press, prosinec 1991.

[24] Stapledon, Olaf, The Starmaker (Tvůrce hvězd), Berkley Books, 1961 (napsáno však bylo pravděpodobně před r. 1937).

[25] Stent, Gunther S., The Coming of the Golden Age: A view of the End of Progress (Příchod Zlatého věku: vyhlídka na konec vývoje), The Natural History Press, 1969.

[26] Swanwick Michael, Vacuum Flowers (Květiny ve vzduchoprázdnu), na pokračování v Isaac Asimov's Science Fiction Magazine, prosinec(?) 1986 - únor 1987. Znovu vydáno v Ace Books, 1988.

[27] Thearling, Kurt, "How We Will Build a Machine that Thinks" (Jak sestrojíme stroj, který myslí), seminář fy Thinking Machines Corporation, 24.-26. srpna 1992. Soukromé sdělení.

[28] Ulam, S., Tribute to John von Neumann (Pocta J. von Neumannovi), Bulletin of the American Mathematical Society, sv. 64, č. 3, část 2, květen 1958, str. 1-49.

[29] Vinge, Vernor, "Bookworm, Run!" (Knihomole, utíkej!), Analog, březen 1966, str. 8-40. Přetištěno v True Names and Other Dangers, Vernor Vinge, Baen Books, 1987.

[30] Vinge, Vernor, "True Names" (Pravá jména), Binary Star Number 5, Dell, 1981. Přetištěno v True Names and Other Dangers, Vernor Vinge, Baen Books, 1987.

[31] Vinge, Vernor, First Word (První slovo), Omni, leden 1983, str.10.

[32] Vinge, Vernor, bude publikováno [:-)].

http://www.transhumanismus.cz/library.php?source=kucera