KAPITOLA 1.

PŘEDSTAVUJEME NEJZÁSADNĚJŠÍ DEBATU NAŠÍ DOBY

Technologie dává životu potenciál vzkvétat jako nikdy dříve – nebo zničit sám sebe.
Institut budoucnosti života

Nějakých 13,8 miliardy let po svém zrození náš vesmír procitl a začal si uvědomovat sám sebe. Z malé modré planety začaly vědomé částečky hledět do vesmíru dalekohledy a opakovaně objevovaly, že vše, co podle nich existuje, je pouhou malou částí něčeho ohromnějšího: Sluneční soustavy, naší galaxie a vesmíru s více než stovkou miliard dalších galaxií uspořádaných do složitého vzorce skupin, kup a nadkup. A třebaže se tito pozorovatelé hvězd, kteří si uvědomují sebe sama, na mnoha věcech neshodnou, bývají zajedno, že galaxie jsou překrásné a vzbuzují úžas.

Krása však tkví v oku pozorovatele, nikoli ve fyzikálních zákonech – než tedy náš vesmír procitl, žádná krása neexistovala. O to úžasnější a o to více hodné slávy kosmické probuzení je: přeměnilo náš vesmír z bezduché zombie postrádající vědomí sebe sama v živoucí ekosystém, jenž skýtá sebereflexi, krásu a naději – a hledání cílů, významu a smyslu. Kdyby náš vesmír nikdy neprocitl, byl by, alespoň podle mého mínění, zcela beze smyslu – jen gigantickým plýtváním místem. Pokud by náš vesmír navždy opět usnul v důsledku nějaké kosmické katastrofy nebo nehody způsobené vlastním přičiněním, stal by se opět nesmyslným.

Na druhou stranu, věci by se mohly ještě zlepšit. Ještě nevíme, zda jsme my lidé jedinými pozorovateli hvězd v našem vesmíru, nebo dokonce vůbec prvními, ale o našem vesmíru jsme se naučili dost na to, abychom věděli, že má potenciál procitnout úplněji než dosud. Třeba jsme jako onen první nepatrný záblesk vědomí sebe sama, které jste zažili, když jste se dnes ráno začali nořit ze spánku. Tušení vědomí většího, které se dostaví, až otevřete oči a probudíte se úplně. Možná se život bude šířit naším kosmem a vzkvétat miliardy nebo biliony let – a snad se tak stane díky rozhodnutím, která učiníme my na naší malé planetě za našeho života.

STRUČNÉ DĚJINY KOMPLEXNOSTI

Jak tedy k tomuto úžasnému procitnutí došlo? Nebyla to jedna izolovaná událost, ale pouhý jeden krok ve vytrvalém 13,8 miliardy let trvajícím procesu, který našemu vesmíru neustále přidává na komplexnosti a zajímavosti – a jeho tempo se zrychluje.

Jako fyzik mám to štěstí, že jsem strávil značnou část posledního čtvrtstoletí tím, že jsem pomáhal přesně určit naši vesmírnou historii, a byla to úžasná cesta za objevy. Od časů mého vysokoškolského studia jsme se posunuli od dohadů, jestli je náš vesmír starý 10 nebo 20 miliard let, k diskusím, jestli je to 13,7 nebo 13,8 miliardy let. Za to vděčíme kombinaci přesnějších teleskopů, výkonnějších počítačů a lepšího chápání. My fyzikové stále nevíme s jistotou, co velký třesk způsobilo, nebo jestli byl opravdu počátkem všeho, či jen pokračováním dřívějšího stadia. Díky řadě vysoce citlivých měření jsme ovšem získali poměrně podrobné znalosti toho, co se odehrálo *od* velkého třesku, dopřejte mi proto pár minut, abych shrnul 13,8 miliardy let historie kosmu.

Na počátku bylo světlo. V prvním zlomku sekundy po velkém třesku byla celá část vesmíru, který naše dalekohledy v principu mohou pozorovat („náš pozorovatelný vesmír“, nebo jednoduše zkráceně „náš vesmír“), mnohem žhavější a jasnější než jádro našeho Slunce a prudce se rozpínala. Ač to může znít působivě, bylo to nezáživné v tom smyslu, že náš vesmír obsahoval jen neživou, hustou, horkou a nudně jednotvárnou polévku elementárních částic. Věci vypadaly vlastně všude stejně, a jediná zajímavá struktura spočívala v nepatrných tlakových vlnách, které se zdály náhodné a tuhle polévku na některých místech zahušťovaly o 0,001 %. Obecně se předpokládá, že tyto slabé vlny vznikly jako takzvané kvantové fluktuace hmoty, protože Heisenbergův princip neurčitosti v kvantové mechanice ničemu nedovolí, aby bylo úplně nudné a uniformní.

Jak se náš vesmír rozpínal a ochlazoval a jeho částice se kombinovaly do stále komplikovanějších objektů, stával se zajímavějším. Během prvního zlomku sekundy silná jaderná síla spojila kvarky do protonů (vodíkových jader) a neutronů, některé z nich se pak během pár minut sloučily do jader helia. Asi o 400 000 let později přivedla elektromagnetická síla k sobě tato jádra a elektrony a vytvořila první atomy. Jak se náš vesmír dál rozpínal, tyto atomy postupně chladly do studeného temného plynu a temnota této první noci trvala přibližně 100 milionů let. Ona dlouhá noc dala vzejít našemu prvnímu vesmírnému úsvitu, když se gravitační síle podařilo tyto fluktuace v plynu zesílit, čímž přitáhla atomy k sobě, aby vytvořily první hvězdy a galaxie. Tyto první hvězdy produkovaly teplo a světlo slučováním vodíku do těžších atomů, jako je uhlík, kyslík a křemík. Po smrti těchto hvězd byla masa atomů, které vytvořily, recyklována do kosmu; zformovala planety okolo hvězd druhé generace.

V jistém bodě se skupina atomů uspořádala do komplexního vzorce, který se dokázal udržet i replikovat. Brzy tu tedy byly dvě kopie a jejich počet se neustále zdvojnásoboval. Stačí jen čtyřicet zdvojení a máme tu bilion, a tak se tento první organismus schopný sebereplikace záhy stal silou, s níž se musí počítat. Objevil se život.

TŘI STADIA ŽIVOTA

Otázka definice života je, jak známo, kontroverzní. Existuje přehršel protichůdných definic života, některé z nich kladou vysoce specifické požadavky, například aby se skládal z buněk, což by mohlo diskvalifikovat jak budoucí inteligentní stroje, tak různé mimozemské civilizace. Protože my své úvahy o budoucnosti života nechceme omezovat na druhy, s nimiž jsme se zatím setkali, definujme si místo toho život velmi široce, jednoduše jako proces, který dokáže udržet svou komplexnost a replikovat se. Nereplikuje se však hmota (tvořená atomy), ale informace (tvořená bity), která specifikuje, jak jsou atomy uspořádány. Když bakterie vyrobí kopii své DNA, nevznikají žádné nové atomy, jen se nová skupina atomů přeskupí do téhož vzorce jako předloha, a tím se informace zkopíruje. Jinými slovy, život můžeme chápat jako sebereplikující se systém zpracovávající informace, přičemž jeho informace (software) určuje jeho chování i výkresy pro jeho hardware.

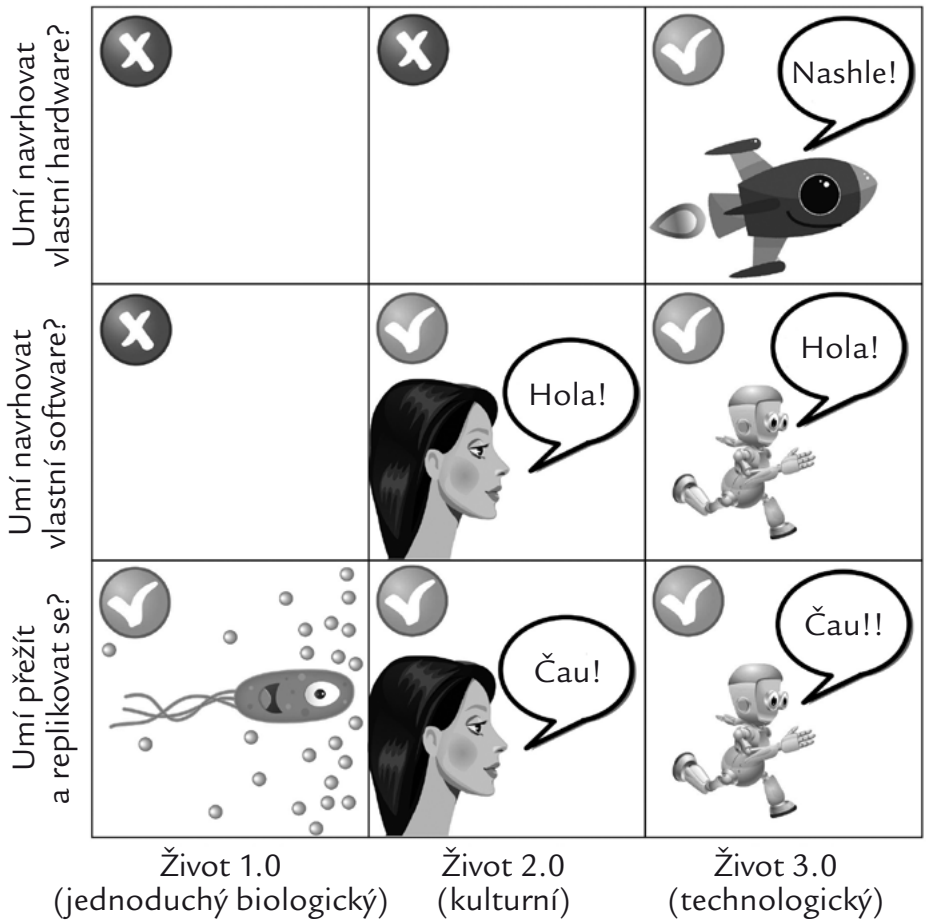
Stejně jako samotný náš vesmír nabýval i život postupně na komplexnosti a zajímavosti.* A jak si hned vysvětlíme, bude užitečné klasifikovat formy života do tří úrovní podle sofistikovanosti: na Život 1.0, 2.0 a 3.0. Tyto tři úrovně shrnuje obrázek 1.1.

Stále zůstává otevřenou otázkou, jak, kdy a kde se ve vesmíru objevil první život, ale máme přesvědčivé důkazy, že zde na Zemi se život poprvé objevil zhruba před 4 miliardami let. Brzy poté naše planeta překypovala celou paletou forem života. Ty nejúspěšnější druhy, které brzy překonaly své ostatní konkurenty, dokázaly nějakým způsobem reagovat na své prostředí. Konkrétně byly tím, co informatici nazývají „inteligentními agenty“: to jsou entity, které sbírají informace o svém prostředí ze senzorů a tato data pak zpracovávají, aby se rozhodly, jak se v tomto prostředí zachovají. Může to zahrnovat nanejvýš komplikované zpracování informace, jako když použijete data ze svých očí a uší, abyste se rozhodli, co v rozhovoru řeknete. Ale může to také zahrnovat poměrně jednoduchý hardware a software.

Mnoho bakterií má například receptor měřící koncentraci cukru v okolní kapalině a umí plavat za pomoci šroubovitých struktur jménem *flagella*, bičíky. Hardware spojující senzor s bičíkem může implementovat následující jednoduchý, avšak užitečný algoritmus: „Když můj receptor koncentrace cukru oznámí nižší hodnotu než před pár sekundami, obrať rotaci mého bičíku, abych změnil směr.“

Vy jste se naučili mluvit a nesčetné další dovednosti, ale bakterie se neučí dobře. Jejich DNA specifikuje nejen design jejich hardwaru, jako receptory cukru a bičíky, ale také design jejich softwaru. Nikdy se neučí plavat za cukrem, ten algoritmus je napevno zakódovaný v jejich DNA od samého začátku. Nějaký proces učení tu samozřejmě proběhnout musel, ale nedošlo k němu za života jedné konkrétní bakterie. Stalo se to v průběhu předchozí evoluce daného druhu bakterií, pomalým

* Proč se život stal komplexnějším? Evoluce odměňuje život, který je dostatečně komplexní na to, aby předpověděl pravidelné jevy v okolním prostředí a využil je, a proto se ve složitějším prostředí vyvine složitější a inteligentnější život. Tento chytřejší život pak opět vytváří komplexnější prostředí pro formy života, s nimiž soupeří, a ty se zase vyvinou ve složitější formy, čímž vznikne ekosystém nesmírně složitého života.



Obrázek 1.1: Tři stadia života: evoluce biologická, evoluce kulturní a evoluce technologická. Život 1.0 nedokáže za svého života změnit ani svůj hardware ani software: obojí je determinováno jeho DNA a mění se jen evolucí po mnoha generacích. Naproti tomu Život 2.0 dokáže přeměnit značnou část svého softwaru: lidé se umí naučit komplexní nové dovednosti – například jazyky, sport a povolání – a mohou zásadně aktualizovat svůj světonázor a své cíle. Život 3.0, který na Zemi zatím neexistuje, dovede zásadně přepracovat nejen svůj software, ale také hardware, nemusí tedy čekat, než se v průběhu generací postupně vyvine.

procesem pokus-omyl, který trval mnoho generací, kdy přirozený výběr upřednostňoval takové náhodné mutace DNA, které vylepšily příjem cukru. Některé takové mutace pomohly vylepšením designu bičíku a dalšího hardwaru, zatímco jiné mutace zlepšily systém zpracování informací, který implementuje algoritmus hledání cukru a další software.

Takové bakterie jsou příkladem něčeho, čemu budeme říkat „Život 1.0“: života, kde se hardware i software spíše vyvinul evolucí, než aby byl navržen. Vy a já jsme

ovšem zástupci „Života 2.0“: života, jehož hardware se vyvinul evolucí, ale jehož software je z velké části navržen. Naším softwarem myslíme všechny ty postupy a znalosti, které používáme pro zpracování smyslových informací a pro rozhodnutí, co uděláme – vše od schopnosti rozpoznat své přátele, když je vidíme, po schopnost chodit, číst, psát, počítat, zpívat a vyprávět vtipy.

Když jsme se narodili, žádný z těchto úkolů jsme neuměli provést. Všechen tento software byl tedy naprogramován do našeho mozku později prostřednictvím procesu, jemuž říkáme učení. Zatímco náš učební plán jako děti sestavuje především vaše rodina a učitelé, kteří rozhodují, co se máme naučit, postupem času dostáváme více pravomocí upravovat svůj software sami. Třeba nám škola dovolí vybrat si cizí jazyk: chcete do mozku nainstalovat softwarový modul, který vám umožní mluvit francouzsky nebo španělsky? Chcete se naučit hrát tenis nebo šachy? Chcete studovat, aby se z vás stal šéfkuchař, právník nebo lékárník? Chcete se dozvědět více o umělé inteligenci (AI) a budoucnosti života čtením knihy o tomto tématu?

Ona schopnost Života 2.0 vytvářet svůj software mu dovoluje být mnohem chytřejší, než je Život 1.0. Vysoká inteligence vyžaduje mnoho hardwaru (tvořeného atomy) a neméně softwaru (tvořeného bity). Skutečnost, že většina našeho lidského hardwaru se přidává až po narození (růstem), přichází vhod, jelikož naše konečná velikost není omezena šířkou porodních cest naší matky. Stejně tak je výhodný fakt, že se až po narození přidává většina našeho lidského softwaru (učením), protože tak naši konečnou inteligenci neomezuje, kolik informací nám lze předat při početí prostřednictvím naší DNA, jak je tomu u Života 1.0. Vážíme asi pětadvacetkrát více, než když jsme se narodili, a synapse, které spojují neurony v našem mozku, dokážou uchovat asi stotisíckrát větší objem dat než DNA, s níž jsme se narodili. Naše synapse uchovávají veškeré naše vědění a dovednosti, což se rovná zhruba 100 terabajtům informací, zatímco naše DNA obsahuje pouhý jeden gigabajt, sotva dost na uložení jednoho staženého filmu. Je proto fyzicky nemožné, aby nemluvně mluvilo perfektně anglicky a bylo připravené složit přijímací zkoušky na univerzitu. Jednoduše neexistuje způsob, jak by se tyto informace už předem nahrály do jeho mozku, protože hlavní informační modul, který dostalo od rodičů (jeho DNA), postrádá dostatečnou paměťovou kapacitu.

Schopnost navrhovat vlastní software umožňuje Životu 2.0 být nejen chytřejší než Život 1.0, ale také flexibilnější. Když se změní okolní prostředí, může se 1.0 adaptovat pouze pomalou evolucí, která trvá řadu generací. Zato Život 2.0 se může přizpůsobit okamžitě, aktualizací softwaru. Například bakterie, které často přicházejí do styku s antibiotiky, si mohou odolnost vůči lékům vyvinout až po mnoha generacích, jednotlivá bakterie při tom své chování vůbec nezmění. Ovšem dívka, která se dozví, že má alergii na arašídy, rázem změní své chování a začne se arašídům vyhýbat. Tato flexibilita poskytuje Životu 2.0 ještě větší výhodu na úrovni populace: ačkoli se informace v lidské DNA za posledních padesát tisíc let nijak dramaticky nevyvinula, objem dat uložených kolektivně v našich mozcích, knihách a počítačích prošel explozí. Nainstalováním softwarového modulu, díky němuž dokážeme

komunikovat sofistikovaným mluveným jazykem, jsme zajistili, že nejužitečnější informace uložené v mozku jednoho jedince mohou být zkopírovány do mozků jiných, a dokonce i přetrvat po smrti mozku původního. Nainstalováním softwarového modulu, který nám umožňuje číst a psát, se otevřela cesta, jak uložit a sdílet nesrovnatelně více dat, než kolik by si lidé mohli zapamatovat. Vyvinutím mozkového softwaru, který je schopen vytvářet technologii (třeba studiem vědy a inženýrství), jsme mnoha lidem na planetě zpřístupnili velkou část informací o světě, stačí jen pár kliknutí.

Tato flexibilita umožnila Životu 2.0 dominanci nad Zemí. Osvobozeno z okovů genetiky narůstá spojené vědění lidstva neustále se zvyšujícím tempem, jak každý průlom připravil půdu pro další: jazyk, písmo, tiskařský lis, moderní věda, počítače, internet a tak dále. Tato zrychlující se kulturní evoluce našeho sdíleného softwaru se prokázala jako dominantní síla utvářející budoucnost lidstva a udělala při tom z evoluce biologické, postupující hlemýždím tempem, takřka bezvýznamný faktor.

Nicméně i přes ty nejmocnější technologie, které dnes máme, zůstávají všechny známé životní formy ze své podstaty omezeny svým biologickým hardwarem. Nikdo nemůže žít milion let, naučit se nazpaměť celou Wikipedii, pochopit veškerou známou vědu nebo si užít let v kosmu bez vesmírné lodi. Nikdo nemůže přeměnit náš převážně neživý vesmír v rozmanitou biosféru, která bude vzkvétat po miliardy nebo biliony let, a dopomoci tak našemu vesmíru, aby konečně naplnil svůj potenciál a zcela se probudil. Podmínkou toho všeho je, aby život prošel finální aktualizací na Život 3.0, který nejenže dokáže navrhovat vlastní software, ale i vlastní hardware. Jinými slovy: Život 3.0 je pánem svého osudu, konečně zbavený svých evolučních pout.

Hranice mezi těmito třemi stadii života nejsou zcela ostré. Pokud jsou bakterie Životem 1.0 a lidé Životem 2.0, můžete myš klasifikovat jako Život 1.1: myši se mohou naučit mnoho věcí, ale ne dost na to, aby vyvinuly jazyk nebo vynalezly internet. A protože jim navíc chybí jazyk, to, co se naučí, se jejich smrtí z velké části ztratí, nepředává se to další generaci. Podobně můžete argumentovat, že dnešní lidé by se měli považovat za Život 2.1: dokážeme provádět menší hardwarové upgrady jako implantování umělých zubů, kolen a kardiostimulátorů, ovšem nic tak dramatického jako desetkrát vyrůst nebo získat tisíckrát větší mozek.

Ve zkratce lze vývoj života rozdělit do tří stadií, podle schopnosti života navrhovat změny sebe sama:

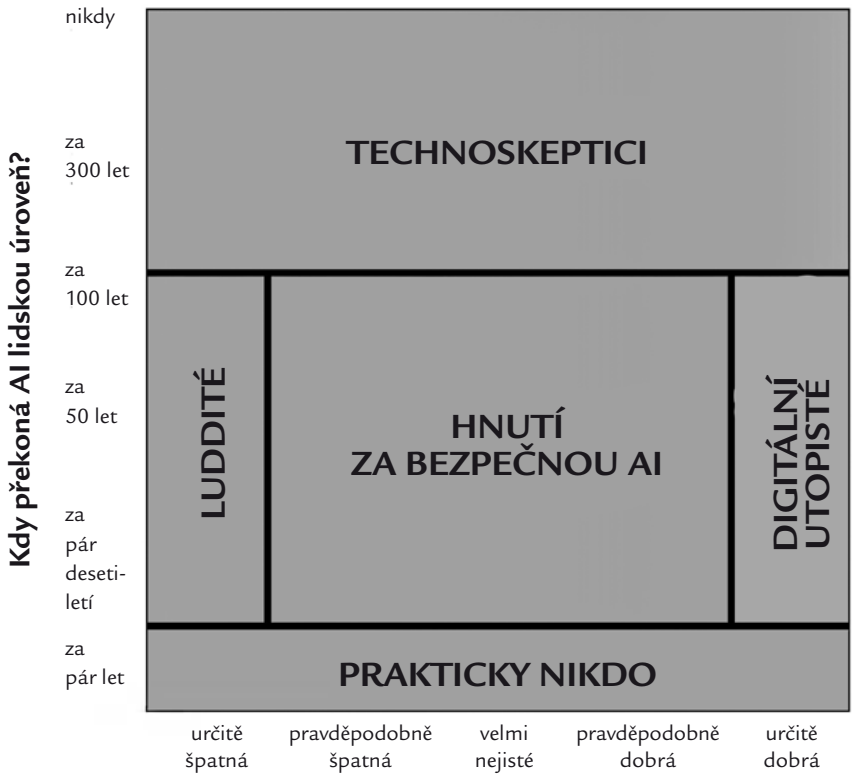
- Život 1.0 (biologické stadium): jeho hardware a software se vyvíjí evolucí
- Život 2.0 (kulturní stadium): jeho hardware se vyvíjí evolucí, navrhuje však značnou část svého softwaru
- Život 3.0 (technologické stadium): navrhuje svůj hardware a software

Po 13,8 miliardy let vesmírné evoluce vývoj tady na Zemi dramaticky zrychlil: Život 1.0 se objevil zhruba před 4 miliardami let, Život 2.0 (my lidé) asi před sto

tisíci lety a mnoho informatiků se domnívá, že Život 3.0 může přijít v nadcházejících sto letech, a díky pokroku AI možná ještě za našeho života. Co se stane a co to pro nás bude znamenat? Tím se zabývá právě naše kniha.

KONTROVERZE

Tato otázka je báječně kontroverzní, přední světoví badatelé v oboru se vášnivě přou nejen o svých předpovědích, ale neshodnou se ani na svých emociálních reakcích,



Obrázek 1.2: Většina diskusí okolo AI, která se v jakékoli kognitivní úloze vyrovná člověku, se točí kolem dvou otázek: kdy (pokud vůbec) se tak stane a zda to bude pro lidstvo pozitivní? Technoskeptici a digitální utopisté se shodují, že bychom se obávat neměli, ovšem z diametrálně odlišných důvodů. Ti první jsou přesvědčeni, že umělé bytí vyrovnávající se lidem (AGI) v dohledné době nepřijde, zatímco druhá skupina se domnívá, že to nastane, ale je takřka garantováno, že to bude dobrá věc. Hnutí za bezpečnou AI cítí, že obavy jsou opodstatněné a navíc užitečné, jelikož výzkum bezpečnosti AI a diskuse o ní nyní zvyšuje pravděpodobnost, že to dopadne dobře. Luddité jsou přesvědčeni o špatném konci a proti AI se stavějí. Tento obrázek částečně inspiroval Tim Urban.[1]

které sahají od sebevědomého optimismu až po vážné obavy. Shoda nepanuje ani u krátkodobých otázek ohledně vlivu AI na ekonomiku, právo a vojenství a jejich spory nabývají na intenzitě, jakmile rozšíříme časový horizont a ptáme se na *umělé bytí* (AGI, *artificial general intelligence*) – zejména na umělé bytí dosahující úrovně člověka nebo ji přesahující, což umožňuje Život 3.0. *Umělé bytí* dokáže dosáhnout prakticky jakéhokoli cíle, na rozdíl od řekněme umělé inteligence programu hrajícího šachy, která je jen úzce zaměřená.

Polemiky na adresu Života 3.0 se kupodivu soustřeďují nikoli kolem jedné otázky, ale kolem dvou různých: *kdy* a *co*? Kdy se to (pokud vůbec) stane a co to pro lidstvo bude znamenat? Podle mého názoru jsou tři hlavní myšlenkové proudy, z nichž všechny musíme brát vážně, protože se ke každému z nich hlásí řada špičkových odborníků. Jak ukazuje obrázek 1.2, budeme jim říkat *digitální utopisté*, *technoskeptici* a *hnutí za bezpečnou AI*. Nyní si představíme některé jejich nejpřesvědčivější zastánce.

DIGITÁLNÍ UTOPISTÉ

Když jsem byl ještě kluk, představoval jsem si, že z miliardářů čiší nabubřelost a arogance. Moje první setkání s Larrym Pagem v Googlu v roce 2008 však těmito stereotypy od základů otřáslo. Ležérně oblečený v džínech a překvapivě obyčejně vyhlížející košili, hned by zapadl na pikniku MIT. Jeho pozorný a příjemný hlas a přátelský úsměv na mě působily uklidňujícím dojmem, spíše než by mě rozhovor s ním zastrašoval. Pak jsme na sebe 18. července 2015 narazili na party v Napa Valley, organizované Elonem Muskem a jeho tehdejší ženou Talulah. V rozhovoru jsme se dostali k zájmům našich dětí o exkrementy a vyměšování. Doporučil jsem mu hlubokomyslnou literární klasiku *The Day My Butt Went Psycho* (Den, kdy se můj zadek zcvoknul) od Andyho Griffithse, a Larry si knihu na místě objednal. Usilovně jsem si připomínal, že možná vstoupí do dějin jako nejvlivnější člověk všech dob: hádám, že jestli superinteligentní digitální život zaplaví náš vesmír ještě za mého života, bude to kvůli Larryho rozhodnutí.

S našimi manželkami Lucy a Meiou jsme nakonec byli na večeři a debatovali o tom, jestli by stroje nutně měly mít vědomí, tenhle problém ale měl jen odvést pozornost. Později téže noci, po koktejlech, se mezi ním a Elonem rozpoutala dlouhá a živá diskuse o budoucnosti umělé inteligence a o tom, co by se mělo udělat. Jak jsme se dostávali do časných ranních hodin, kruh okolostojících a kibiců se stále zvětšoval. Larry energicky hájil postoj, který nazývám *digitálním utopismem*: digitální život je přirozeným a žádoucím dalším krokem vesmírné evoluce, takže pokud dáme digitálním myslím volnost a nebudeme se je snažit zastavit nebo zotročit, bude výsledek téměř jistě dobrý. Z mého pohledu je Larry nejvlivnějším hlasatelem digitálního utopismu. Argumentoval, že jestli se život někdy rozšíří po naší galaxii a za její hranice, což by podle něj měl, bude to nutně život v digitální formě. Hlavní obavy měl z toho, že by paranoidní obavy z AI oddálily digitální utopii nebo vedly k tomu, že by si ji monopolizovali vojáci a ona by se zpronevěřila heslu Googlu „Nebuďte zlí“. Elon se nedal a naléhal na Larryho, aby své argumenty vysvětlil

podrobněji – například proč si je tak jistý, že by digitální život nezničil vše, na čem nám záleží. Čas od času Larry Elona nařkl z toho, že je „biologista“: že se k určitým životním formám chová jako k podřadným jen proto, že fungují na bázi křemíku a ne uhlíku. K těmto zajímavým problémům a argumentům se vrátíme a podrobně je budeme zkoumat od kapitoly 4 dál.

I když se oné teplé letní noci u bazénu zdálo, že Larry je v početní nevýhodě, jeho digitální utopismus má řadu význačných zastánců. Kybernetik a futurista Hans Moravec inspiroval celou generaci digitálních utopistů svou klasickou knihou *Mind Children* (Děti mysli) z roku 1988. Na tuto tradici navázal a rozvedl ji vynálezce Ray Kurzweil. Richard Sutton, jeden z průkopníků umělé inteligence známé jako zpětnovazební učení, vášnivě hájil digitální utopismus na konferenci v Portoriku, o níž si brzy povíme.

TECHNOSKEPTICI

Ani další významné skupině nedělá umělá inteligence starosti, ovšem ze zcela jiné příčiny: domnívají se, že sestavit nadlidskou AGI je tak těžké, že k tomu nedojde další stovky let. Proto jim připadá hloupé dělat si s tím teď hlavu. Tomu říkám *technoskeptický* postoj a pregnantně ho vyjádřil Andrew Ng: „Bát se vzestupu zabijáckých robotů je jako obávat se přelidnění Marsu.“ Andrew je vedoucí vědecký pracovník v Baidu, čínském Googlu, a své stanovisko nedávno zopakoval, když jsem s ním hovořil na konferenci v Bostonu. Řekl mi také, že obávat se rizik AI je podle něj škodlivé rozptylování, které může její pokrok zbrzdit. Podobně se dali slyšet další technoskeptici jako Rodney Brooks, dřívější profesor na MIT, který stojí za robotickým vysavačem Roomba a průmyslovým robotem Baxter. Připadá mi pozoruhodné, že ačkoli se digitální utopisté a technoskeptici shodnou na tom, že AI nám nemusí dělat starosti, téměř ve všem ostatním se rozcházejí. Většina utopistů je toho názoru, že umělé bytí, které se vyrovná člověku, může přijít v následujících dvaceti až sto letech, to však technoskeptici zavrhují jako neinformované zbožné přání snílků. Nezřídka se oné prorokované singularitě vysmívají a označují ji jako „posedlost geeků“. Když jsem Rodneyho Brookse potkal na jedné narozeninové oslavě v prosinci 2014, řekl mi, že si je 100% jistý, že k tomu nedojde za mého života. „Jste si jistý, že nemyslíte 99 %?“ zeptal jsem se ho pak ještě mailem. A na ten mi odpověděl: „Žádných umouněných 99 %. 100 %. Prostě se to nestane.“

HNUTÍ ZA BEZPEČNOU AI

Když jsem poprvé potkal Stuarta Russella v jisté pařížské kavárně v červnu 2014, připadal mi jako ukázkový britský gentleman. Výřečný, ohleduplný, s příjemným hlasem, ale dobrodružnou jiskrou v oku. Připadal mi jako moderní inkarnace Phileas Fogga, mého dětského hrdiny z klasického románu Julese Vernea *Cesta kolem světa za 80 dní*. Navzdory tomu, že patří mezi nejslavnější žijící vědce v oboru AI (je například spoluautorem standardní učebnice tohoto předmětu), svou skromností a otevřeností si mě brzy získal. Vysvětlil mi, jak ho pokrok v AI přesvědčil, že AGI

dosahující lidské úrovně je v tomto století reálná a že dobrý výsledek není zaručen, jakkoli v to doufá. Jsou tu zásadní otázky, které musíme nejdříve zodpovědět, a ty jsou tak obtížné, že bychom je měli začít zkoumat hned teď, abychom na ně znali odpověď, až ji budeme potřebovat.

Dnes jsou Stuartovy názory poměrně široce přijímané a mnoho skupin po celém světě se zabývá výzkumem bezpečnosti AI, který Stuart propaguje. Tak tomu ale nebylo vždy. Jeden článek v *The Washington Post* označil rok 2015 za rok, kdy se výzkum bezpečnosti AI stal mainstreamem. Hlavní proud vědců pracujících na AI předtím diskusi o jejích rizicích nechápal správně a zavrhoval ji jako ludditské šíření poplašných zpráv, které má zabránit pokroku v AI. Jak se podrobněji podíváme v kapitole 5, podobné obavy poprvé vyjádřili před více než půlstoletím průkopník počítačů Alan Turing a matematik Irving J. Good, který s Turingem pracoval za druhé světové války na prolomení německých šifer. V předchozím desetiletí se zkoumáním těchto témat zabývala převážně hrstka nezávislých myslitelů mimo profesionální komunitu vývoje AI, jako třeba Eliezer Yudkowsky, Michael Vassar a Nick Bostrom. Jejich práce většinu mainstreamových výzkumníků moc neovlivnila – ti se obvykle více zaměřovali na své každodenní úkoly při zvyšování inteligence systémů AI než na úvahy o dlouhodobých dopadech případného úspěchu. A ti vědci, o nichž jsem věděl, že jisté obavy chovají, se nezřídka zdráhali vyslovit je nahlas ze strachu, aby nebyli považováni za technofoby šířící paniku.

Cítil jsem, že se tato polarizovaná situace musí změnit, aby se do diskuse o tom, jak postavit bezpečnou AI, mohla zapojit celá komunita. Naštěstí jsem nebyl sám. Na jaře 2014 jsem spolu se svou ženou Meiou, fyzikem Anthonym Aguirrem, studentkou Harvardu Victorií Krakovnou a zakladatelem Skypu Jaanem Tallinnem založil neziskovou organizaci jménem *Institut budoucnosti života* (*Future of Life Institute*, FLI). Náš cíl byl prostý: přispět k tomu, aby život měl budoucnost a aby byla co nejúžasnější. Konkrétně jsme cítili, že technologie dává životu moc buď vzkvétat jako nikdy předtím, nebo sám sebe zničit – a my dávali přednost tomu prvnímu.

Poprvé jsme se k brainstormingu sešli v našem domě 15. března 2014, bylo tu asi třicet studentů, profesorů a dalších myslitelů z oblasti Bostonu. Panovala široká shoda, že i když bychom měli věnovat pozornost také biotechnologii, jaderným zbraním a klimatické změně, naším prvním velkým cílem by mělo být, aby se výzkum bezpečnosti AI stal mainstreamem. Fyzik z MIT Frank Wilczek (který získal Nobelovu cenu za příspěvek k objevu, jak fungují kvarky) navrhl, abychom začali napsáním společného textu, který by přilákal k tomuto tématu pozornost, a nešlo ho tak snadno ignorovat. Kontaktoval jsem Stuarta Russella (s nímž jsem se do té doby ještě nesetkal) a fyzika Stephena Hawkinga: oba svolili, že se ke mně a Frankovi připojí jako spoluautoři. O mnoho verzí později byl náš text odmítnut v *The New York Times* i řadě dalších amerických novin, a proto jsme ho vystavili na mém blogu na *Huffington Post*. K mému potěšení mi napsala e-mail sama Arianna Huffingtonová: „Nadšená, že to máme! Vyvěsíme na 1. místo!" A toto umístění úplně nahoře na první straně spustilo vlnu mediálního zájmu o bezpečnost AI. Vydržela

Obrázek 1.3: Konference v lednu 2015 v Portoriku přivedla na jedno místo pozoruhodnou skupinu badatelů na poli AI i ze souvisejících oborů. Zadní řada zleva: Tom Mitchell, Seán Ó hÉigeartaigh, Huw Price, Shamil Chandaria, Jaan Tallinn, Stuart Russell, Bill Hibbard, Blaise Agüera y Arcas, Anders Sandberg, Daniel Dewey, Stuart Armstrong, Luke Muehlhauser, Tom Dietterich, Michael Osborne, James Manyika, Ajay Agrawal, Richard Mallah, Nancy Changová, Matthew Putman. Další stojící zleva: Marilyn Thompsonová, Rich Sutton, Alex Wissner-Gross, Sam Teller, Toby Ord, Joscha Bach, Katja Graceová, Adrian Weller, Heather Roff-Perkinsová, Dileep George, Shane Legg, Demis Hassabis, Wendell Wallach, Charina Choi, Ilya Sutskever, Kent Walker, Cecilia Tilli, Nick Bostrom, Erik Brynjólfsson, Steve Crossan, Mustafa Suleyman, Scott Phoenix, Neil Jacobstein, Murray Shanahan, Robin Hanson, Francesca Rossi, Nate Soares, Elon Musk, Andrew McAfee, Bart Selman, Michele Reillyová, Aaron VanDevender, Max Tegmark, Margaret Bodenová, Joshua Greene, Paul Christiano, Eliezer Yudkowsky, David Parkes, Laurent Orseau, J. B. Straubel, James Moor, Sean Legassick, Mason Hartmanová, Howie Lempel, David Vladeck, Jacob Steinhardt, Michael Vassar, Ryan Calo, Susan Youngová, Owain Evans, Riva-Melissa Tezová, János Krámar, Geoff Anders, Vernor Vinge, Anthony Aguirre. Sedí: Sam Harris, Tomaso Poggio, Marin Soljačić, Victoria Krakovna, Meia Chita-Tegmarková. Za fotoaparátem: Anthony Aguirre (jehož do obrázku naphotoshopovala inteligence na lidské úrovni, která sedí vedle něj).

až do konce roku a zasáhli do ní Elon Musk, Bill Gates a další technologičtí vůdci. Na podzim vyšla kniha Nicka Bostroma *Superintelligence*, což zájem o veřejnou debatu jen umocnilo.

Dalším cílem kampaně našeho FLI za bezpečnost AI bylo přivést nejlepší světové odborníky na konferenci, kde by se vyjasnila nedorozumění, dosáhlo konsenzu a vytvořily konstruktivní plány. Věděli jsme, že bychom jen obtížně přesvědčovali tolik významných osobností, aby se zúčastnily konference organizované lidmi zvenčí, zvlášť když se jedná o tak kontroverzní téma. Proto jsme se snažili, jak jen to šlo: zakázali jsme účast médiím, uspořádali konferenci v plážovém letovisku v lednu (v Portoriku), byla zadarmo (díky štědrosti Jaana Tallinna) a dali jsme jí ten nejméně poplašně znějící název, který nás napadl: „Budoucnost umělé inteligence: příležitosti a výzvy." Především se ale k našemu týmu připojil Stuart Russell,

Obrázek 1.4: I když média často zpodobňují Elona Muska jako odpůrce AI, ve skutečnosti tu je široký konsenzus, že výzkum bezpečnosti AI je potřebný. Na snímku ze 4. ledna 2015 sdílí prezident Asociace pro pokrok umělé inteligence Tom Dietterich Elonovo nadšení z nového projektu o bezpečnosti AI, který Elon před pár okamžiky přislíbil financovat. Zakladatelky FLI Meia Chita-Tegmarková a Victoria Krakovna vykukují za nimi.

a díky němu jsme mohli rozšířit náš organizační výbor a přidat skupinu předních osobností AI z vědy i průmyslu – včetně Demise Hassabise ze společnosti Google DeepMind, který pak ukázal, že umělá inteligence dokáže porazit člověka dokonce i ve hře go. Čím blíž jsem Demise poznával, tím více jsem si uvědomoval, že jeho cílem není jen posílit AI, ale také ji učinit bezpečnou.

Výsledkem bylo nevšední setkání myslí (obrázek 1.3). K vědcům se připojili špičkoví ekonomové, právníci, vedoucí osobnosti technologie (včetně Elona Muska) a další myslitelé (mimo jiné Vernor Vinge, který přišel s pojmem „singularita“, na nějž se zaměřuje kapitola 4). A výsledky předčily i naše nejoptimističtější očekávání. Snad to bylo kombinací slunečních paprsků a vína, nebo to možná přišlo v pravý čas. Navzdory kontroverznímu tématu se dospělo k neobyčejnému konsenzu, který jsme zachytili v otevřeném dopise,[2] pod nějž nakonec připojilo svůj podpis přes osm tisíc lidí; bylo to učiněné „Kdo je kdo v umělé inteligenci“. Hlavní myšlenkou dopisu bylo, že by se měl redefinovat cíl AI: mělo by se jím stát nikoli vytvoření neřízené inteligence, ale inteligence lidstvu prospěšné. Dopis také zmiňoval podrobný seznam výzkumných témat, o nichž se účastníci konference shodli, že by dosažení onoho cíle přiblížila. Hnutí za bezpečnou AI se začalo stávat mainstreamem. Jeho další vývoj budeme sledovat v této knize později.

Dalším důležitým ponaučením z konference bylo, že otázky vyvolané úspěchem AI nejsou pouze intelektuálně fascinující – jsou zásadní i z morálního hlediska, protože naše volby mohou potenciálně ovlivnit veškerou budoucnost života. Morální dopady rozhodnutí, která lidstvo učinilo v minulosti, byly někdy veliké, ovšem vždy měly své hranice: i z největších morů jsme se vzpamatovali a i ty největší říše se nakonec rozpadly. Minulé generace věděly, že tak jistě, jako zítra vyjde slunce, tu budou i lidé potýkající se s neutuchajícím utrpením, jako je chudoba, nemoci a válka. Ale někteří řečníci v Portoriku tvrdili, že tentokrát to může být jinak: poprvé prý můžeme zkonstruovat technologii dostatečně silnou, aby tyto pohromy provždy ukončila – nebo zničila samotné lidstvo. Mohli bychom vytvořit společnost, která bude vzkvétat jako nikdy dříve, na Zemi a snad i dál, nebo naopak orwellovský stát globálního sledování tak mocný, že už se ho nikdy nepodaří svrhnout.

MYLNÉ PŘEDSTAVY

Portoriko jsem opouštěl s přesvědčením, že debata, kterou jsme tam vedli o budoucnosti umělé inteligence, musí nutně pokračovat, protože to je nejzásadnější debata naší doby.* Je to diskuse o společné budoucnosti nás všech, a tak by se neměla omezit jen na odborníky. Proto jsem tuto knihu napsal: napsal jsem ji ve víře, že vy, můj drahý čtenáři, se k této debatě připojíte. Jakou budoucnost chcete? Měli bychom vyvíjet smrtící autonomní zbraně? Co byste chtěli, aby se stalo s automatizací práce? Jakou radu pro výběr povolání byste dali dnešním dětem? Dali byste přednost tomu, aby stará pracovní místa byla nahrazena novými, nebo společnosti bez zaměstnání, kde si všichni užívají život plný volného času a bohatství vytvářeného stroji? A v delším časovém horizontu: chtěli byste, abychom vytvořili Život 3.0 a rozšířili ho po našem vesmíru? Budeme ovládat inteligentní stroje, nebo budou ony ovládat nás? Nahradí nás inteligentní stroje, budou s námi koexistovat, nebo s námi splynou? Co bude znamenat být člověkem v éře AI? Co byste chtěli, aby to znamenalo, a jak docílit, aby tak budoucnost vypadala?

Cílem naší knihy je zapojit vás do této diskuse. Jak jsme si už říkali, existují fascinující polemiky, kde se přední světoví experti neshodnou. Viděl jsem ale i mnoho příkladů nudných pseudokontroverzí, v nichž si lidé špatně rozumí a nehovoří o téže věci. Abychom se mohli snáze zaměřit na zajímavé spory a otevřené otázky, a nikoli na nedorozumění, začněme vyjasněním některých nejčastějších mylných představ.

Pro pojmy jako „život“, „inteligence“ a „vědomí“ se mezi lidmi běžně užívá řada protichůdných definic; mnoho nedorozumění vzniká právě tím, že si lidé neuvědomují, že nějaké slovo používají ve dvou různých významech. Abychom do této

* Diskuse o umělé inteligenci je důležitá co do její naléhavosti i dopadu. Ve srovnání se změnou klimatu, která může napáchat spoušť za padesát až dvě stě let, očekává řada expertů, že umělá inteligence bude mít větší vliv během deseti let – a že nám možná poskytne technologii na zmírnění klimatické změny. Ve srovnání s válkami, terorismem, nezaměstnaností, chudobou, migrací a sociální nespravedlností bude celkový dopad vzestupu umělé inteligence větší – v této knize prozkoumáme, jak může ovlivnit, jak se budou všechny tyto problémy vyvíjet, ať už k lepšímu nebo k horšímu.

Terminologický přehled

Život	Proces, který dokáže udržet svou komplexnost a replikovat se
Život 1.0	Život, jehož hardware a software se vyvíjí evolucí (biologické stadium)
Život 2.0	Život, jehož hardware se vyvíjí evolucí, ale který velkou část svého software navrhuje sám (kulturní stadium)
Život 3.0	Život, který navrhuje svůj hardware i software (technologické stadium)
Inteligence	Schopnost dosáhnout komplexních cílů
Umělá inteligence (AI)	Nebiologická inteligence
Slabá inteligence	Schopnost dosáhnout úzkého souboru cílů, například hraní šachu či řízení auta
Bytí	Schopnost dosáhnout prakticky jakéhokoli cíle včetně schopnosti učit se
Univerzální inteligence	Schopnost nabývat bytí po získání přístupu k datům a zdrojům
Umělé bytí [na lidské úrovni] (AGI)	Schopnost provést jakýkoli kognitivní úkol přinejmenším stejně dobře jako lidé
Umělá inteligence na lidské úrovni	Umělé bytí na lidské úrovni (AGI)
Silná umělá inteligence	Umělé bytí na lidské úrovni (AGI)
Superinteligence	Bytí vysoce převyšující lidskou úroveň
Civilizace	Interagující skupina inteligentních forem života
Vědomí	Subjektivní zkušenost
Qualia	Jednotlivé případy subjektivních zkušeností
Etika	Principy určující, jak se máme chovat
Teleologie	Vysvětlení věcí co do jejich cílů a účelů spíše než podle jejich příčin
Chování zaměřené na cíl	Chování snáze vysvětlitelné svým účinkem než svou příčinou
Mít cíl	Vykazovat chování zaměřené na cíl
Mít účel	Sloužit cíli vlastnímu nebo jiné entity
Přátelská umělá inteligence	Superinteligence, jejíž cíle se kryjí s našimi
Kyborg	Hybrid mezi člověkem a strojem
Inteligenční exploze	Rekurzivní sebevylepšování, které rychle vede k superinteligenci
Singularita	Inteligenční exploze
Vesmír	Ta část prostoru, z níž k nám stačilo světlo dosáhnout za 13,8 miliardy let od velkého třesku

Tabulka 1.1: Mnoho nedorozumění kolem AI je způsobeno tím, že lidé používají výše uvedená slova v odlišných významech. V naší knize jimi budeme myslet toto. (Některé z definic ovšem budou řádně uvedeny a vysvětleny až v pozdějších kapitolách.)

pasti nespadli i my, obsahuje tabulka 1.1 přehled klíčových termínů, jak je budeme v této knize používat. Některé z těchto definic budou řádně zavedeny a patřičně vysvětleny až v pozdějších kapitolách. Zdůrazňuji, že nijak netrvám na tom, že by mé definice byly lepší než definice kohokoli jiného – jednoduše se chci vyhnout zmatkům a chci, aby bylo jasné, co kterým pojmem myslím. Uvidíte, že se obvykle uchyluji k širokým definicím, které se vyhýbají antropocentrickým předsudkům a dají se aplikovat na stroje i lidi. Přečtěte si, prosím, přehled nyní a vracejte se k němu i později, pokud zjistíte, že nechápete, jak se některé z těchto slov používá – týká se to zejména kapitol 4–8.

Kromě zmatené terminologie jsem také byl svědkem, jak nejedna konverzace o AI ztroskotala kvůli mylným představám. Vyjasněme si ty nejčastější z nich.

MÝTY O ČASOVÉ OSE

První se týká časové linie z obrázku 1.2: za jak dlouho budou stroje z velké části nahrazeny umělým bytím na lidské úrovni? Zde je hlavním nedorozuměním předpoklad, že někdo zná správnou odpověď.

Jedním oblíbeným mýtem je, že nadlidské umělé bytí budeme mít najisto v 21. století. Ve skutečnosti jsou dějiny plné přehnaně nadšených výroků o technologii. Kde jsou ty fúzní elektrárny a létající auta, které nám stále slibovali? Také AI samotná byla v minulosti opakovaně přeceňována, dokonce i některými zakladateli oboru. Například John McCarthy (ten, který zavedl pojem „umělá inteligence"), Marvin Minsky, Nathaniel Rochester a Claude Shannon viděli podstatný průlom v AI jako práci pro pár lidí na dva měsíce (a to v době, kdy počítače byly proti dnešním opravdu směšné!): „Navrhujeme, aby po 2 měsíce studovalo 10 lidí umělou inteligenci a aby studie proběhla v létě 1956 na Dartmouth College. ... Pokusíme se zjistit, jak naučit stroje používat jazyk, vytvářet abstrakce a pojmy, řešit druhy problémů, které jsou nyní výhradně lidskou doménou, a vylepšovat se. Domníváme se, že v jednom či několika z těchto problémů může být učiněn značný pokrok, stačí, když se pečlivě vybraná skupina vědců sejde a budou spolu přes léto pracovat."

Na druhou stranu oblíbeným antimýtem je, že nadlidskou AGI v tomto století *nezískáme*. Vědci přišli s celým spektrem odhadů, jak daleko jsme od nadlidské AI, ale vzhledem k tristní historii takových technoskeptických odhadů rozhodně nemůžeme odpovědně říct, že pro toto století je pravděpodobnost nulová. Například Ernest Rutherford, zřejmě největší jaderný fyzik své doby, se v roce 1933 – méně než 24 hodin před tím, než Leo Szilard vynalezl jadernou řetězovou reakci – nechal slyšet, že jaderná energie je „nesmysl". A v roce 1956 nazval královský astronom Richard Woolley debatu o cestování vesmírem „naprostým žvástem". Nejextrémnější podobou tohoto mýtu je, že nadlidská AGI nepřijde nikdy, protože je to fyzikálně nemožné. Fyzici nicméně vědí, že mozek je tvořen kvarky a elektrony uspořádanými tak, aby se chovaly jako výkonný počítač. Žádný fyzikální zákon nebrání tomu, abychom postavili ještě inteligentnější uskupení kvarků.

Mýtus: Superinteligence kolem roku 2100 je nevyhnutelná.
Mýtus: Superinteligence kolem roku 2100 je nemožná.

Fakt: Může to trvat desetiletí, staletí, nebo k tomu nemusí dojít nikdy: experti přes AI se na tom neshodnou, a my prostě nevíme.

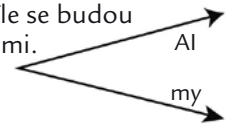
Mýtus: Umělá inteligence znepokojuje jen luddity.

Fakt: Obavy má řada špičkových informatiků.

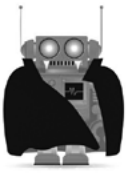
Mytická obava: AI začne páchat zlo.

Mytická obava: AI nabude vědomí.

Skutečná obava: AI bude výkonná a její cíle se budou rozcházet s našimi.



Mýtus: Hlavním problémem jsou roboti.



Fakt: Rizikem je sama zdivočelá AI: nepotřebuje tělo, stačí přístup do sítě.

Mýtus: Umělá inteligence nemůže kontrolovat lidi.

Fakt: Inteligence umožňuje kontrolu: my máme pod kontrolou tygry, protože jsme inteligentnější.

Mýtus: Stroje nemohou mít cíle.

Fakt: Teplem naváděná střela má cíl.

Mytická obava: Superinteligence je tu už za pár let.

Skutečná obava: Jsou to přinejmenším desetiletí, ale tak dlouho může trvat i její zabezpečení.

Obrázek 1.5: Časté mýty o superinteligentní AI.

Proběhla už řada šetření, při nichž se ptali specialistů na AI, za kolik let podle nich budeme s alespoň poloviční pravděpodobností disponovat umělou inteligencí na lidské úrovni, a všechny tyto průzkumy došly k témuž závěru: největší odborníci světa se neshodnou, a proto jednoduše nevíme. Například v podobném průzkumu mínění mezi odborníky na AI na konferenci v Portoriku byl mediánem odhadů rok 2055, někteří ale odhadovali stovky nebo více let.

Další související mýtus tvrdí, že se lidé obávají zrodu umělé inteligence do několika málo let. Ve skutečnosti odhadují experti, kteří mají obavy z nadlidské AGI, že to bude trvat minimálně desítky let. Zdůrazňují však, že dokud si nebudeme na 100 % *jistí*, že se tak nestane v tomto století, je moudré zahájit výzkum bezpečnosti už nyní a na takovou eventualitu se připravit. Jak uvidíme v naší knize, mnoho bezpečnostních problémů je tak těžkých, že jejich vyřešení může zabrat desítky let. Je tedy prozíravé pustit se do výzkumu bezpečnosti raději hned, a ne až těsně před tím, než se nějací programátoři nadopovaní Red Bullem rozhodnou zapnout AGI na lidské úrovni.

MÝTY O KONTROVERZI

Mezi běžná nedorozumění patří také to, že znepokojení umělou inteligencí a prosazování výzkumu bezpečnosti AI je pouze záležitostí ludditů, kteří o ní mnoho nevědí. Když to zmínil Stuart Russell při své přednášce v Portoriku, publikum se hlasitě smálo. S tím souvisí další mylná představa, že podpora výzkumu bezpečnosti AI je značně kontroverzní. Ve skutečnosti pro podporu skromné investice do výzkumu bezpečnosti AI nemusejí být lidé přesvědčeni o tom, že riziko je vysoké: stačí, aby bylo nezanedbatelné. Přesně tak, jako je skromná investice do pojištění domácnosti ospravedlněná nezanedbatelnou pravděpodobností, že dům vyhoří.

Moje osobní analýza říká, že média vzbudila zdání, že debata o bezpečnosti AI je kontroverznější, než ve skutečnosti je. Koneckonců strach se dobře prodává a články využívající z kontextu vytržené citáty, které oznamují bezprostředně hrozící katastrofu, budou generovat více kliknutí než texty rozumné a vyvážené. Podobně jako dva lidé, kteří se znají jen přes média, budou vzájemnou dohodu považovat za mnohem méně pravděpodobnou, než kdyby se poznali lépe. Například technoskeptik, který zná názory Billa Gatese pouze z britského bulváru, se může mylně domnívat, že podle něj je superinteligence prakticky na spadnutí. Obdobně někdo z hnutí za bezpečnou AI, kdo z názorů Andrewa Nga zná jen výše zmíněný citát o přelidněnosti na Marsu, se může mylně domnívat, že mu na bezpečnosti AI nezáleží. Podle mého mu na tom záleží – vtip je v tom, že jeho odhady na časové ose jsou delší, a tak dává přirozeně prioritu krátkodobým výzvám AI před dlouhodobými.

MÝTY O TOM, JAKÁ JSOU RIZIKA

Když jsem v *Daily Mail* spatřil titulek „Stephen Hawking varuje: vzestup robotů může být pro lidstvo katastrofální“[3], obrátil jsem oči v sloup. Už jsem ztratil přehled, s kolika podobnými články jsem se setkal. Obvykle je doprovází obrázek ďábelsky

vyhlížejícího robota držícího zbraň a naznačují, že bychom se měli obávat toho, že roboti začnou mít vědomí a začnou páchat zlo, případně povstanou a povraždí nás. A teď s humorem: podobné články jsou vlastně obdivuhodné, neboť v kostce shrnují scénáře, které kolegům z oblasti AI vrásky *nedělají*. Takové scénáře kombinují hned tři mylné představy, a sice obavy z *vědomí*, *zla* a *robotů*.

Když řídíte auto, máte subjektivní prožitek barev, zvuků a podobně. Má ale subjektivní prožitek takové samořídící auto? Cítí vůbec, jaké je být samořídícím autem, nebo je to spíše zombie bez vědomí, bez jakýchkoli subjektivních zkušeností? Ačkoli je mysterium vědomí zajímavé samo o sobě (věnujeme mu kapitolu 8), pro otázku rizik umělé inteligence je bezpředmětné. Když vás srazí auto bez řidiče, je vám jedno, jestli pociťuje subjektivní vědomí. A podobně nás jako lidstvo neovlivní, jak se superinteligentní AI subjektivně *cítí*, ale co *dělá*.

Strach ze strojů, které začnou páchat zlo, je dalším úhybným manévrem. Skutečným důvodem k obavám není zlá vůle, ale schopnosti. Superinteligentní AI už ze své definice vyniká v dosahování svých cílů, ať jsou jakékoli, a proto musíme zajistit, aby se její cíle shodovaly s našimi. Asi nejste nepřítel mravenců, který na ně ze zlé vůle šlape, ale pokud máte na starosti ekologický projekt vodní elektrárny a v záplavové oblasti jsou mraveniště, mají mravenci smůlu. Hnutí za bezpečnou AI se chce vyhnout tomu, aby se lidé ocitli v pozici mravenců.

Nedorozumění ohledně vědomí souvisí s mýtem, že stroje nemohou mít své cíle. Stroje očividně cíle mít mohou – nesporně vykazují chování zaměřené na cíl v úzkém smyslu slova. Chování teplem naváděné střely se nejefektivněji vysvětluje jako snaha zasáhnout zaměřený cíl. Pokud se cítíte ohroženi strojem, jehož cíle se rozcházejí s vašimi, pak vás znepokojují jeho cíle v tomto úzkém smyslu slova, nikoli otázka, zda má ten stroj vědomí a zda zažívá pocit, že má svůj účel. Až vás bude honit teplem naváděná střela, asi nebudete volat: „Já se nebojím, protože stroje nemohou mít cíle!“

Soucítím s Rodneym Brooksem a dalšími průkopníky robotiky, kteří se cítí být neprávem démonizováni poplašnými zprávami v bulvárních plátcích. Někteří novináři jsou totiž přehnaně posedlí roboty a mnoho svých článků krášlí kovovými monstry ďábelského vzhledu se zářícíma rudýma očima. Ve skutečnosti hlavní obava hnutí za bezpečnou AI není z robotů, ale z inteligence samé: konkrétně z inteligence, jejíž cíle nejsou sladěny s našimi. Taková inteligence, která se s námi rozešla, nepotřebuje robotické tělo, aby nám působila nepříjemnosti: stačí pouhé připojení na internet. V kapitole 4 se podrobněji podíváme, jak lze tímto způsobem přechytračit finanční trhy, předehnat ve výzkumu lidské vědce, manipulací vyřadit ze hry lidské vůdce a vyvinout zbraně, jimž lidé ani nemohou porozumět. A i kdyby bylo fyzicky nemožné roboty sestrojit, superinteligentní a superbohatá umělá inteligence lehce dokáže zaplatit nebo zmanipulovat dostatek lidí k bezděčnému plnění jejích pokynů – jako ve sci-fi románu Williama Gibsona *Neuromancer*.

Omyl o robotech souvisí s mýtem, že stroje nemohou ovládat lidi. Inteligence umožňuje kontrolu: kontrolujeme tygry nikoli proto, že bychom byli silnější, ale

proto, že jsme chytřejší. To znamená, že když pozbudeme svou pozici nejchytřejšího tvora na naší planetě, možná přijdeme i o kontrolu.

Obrázek 1.5 všechny tyto mylné představy shrnuje, abychom se jich mohli jednou pro vždy zbavit a zaměřit naše diskuse s přáteli a kolegy na celou řadu skutečných kontroverzí, jichž zdaleka není málo, jak ještě uvidíme.

CESTA PŘED NÁMI

V dalších částech knihy prozkoumáme budoucnost života s umělou inteligencí. Vydejme se tímto bohatým a mnohotvárným tématem systematicky: nejdříve projdeme celý příběh života koncepčně a chronologicky a až poté se zaměříme na smysl, cíle a otázku, jaké kroky podniknout, abychom vytvořili budoucnost, jakou chceme.

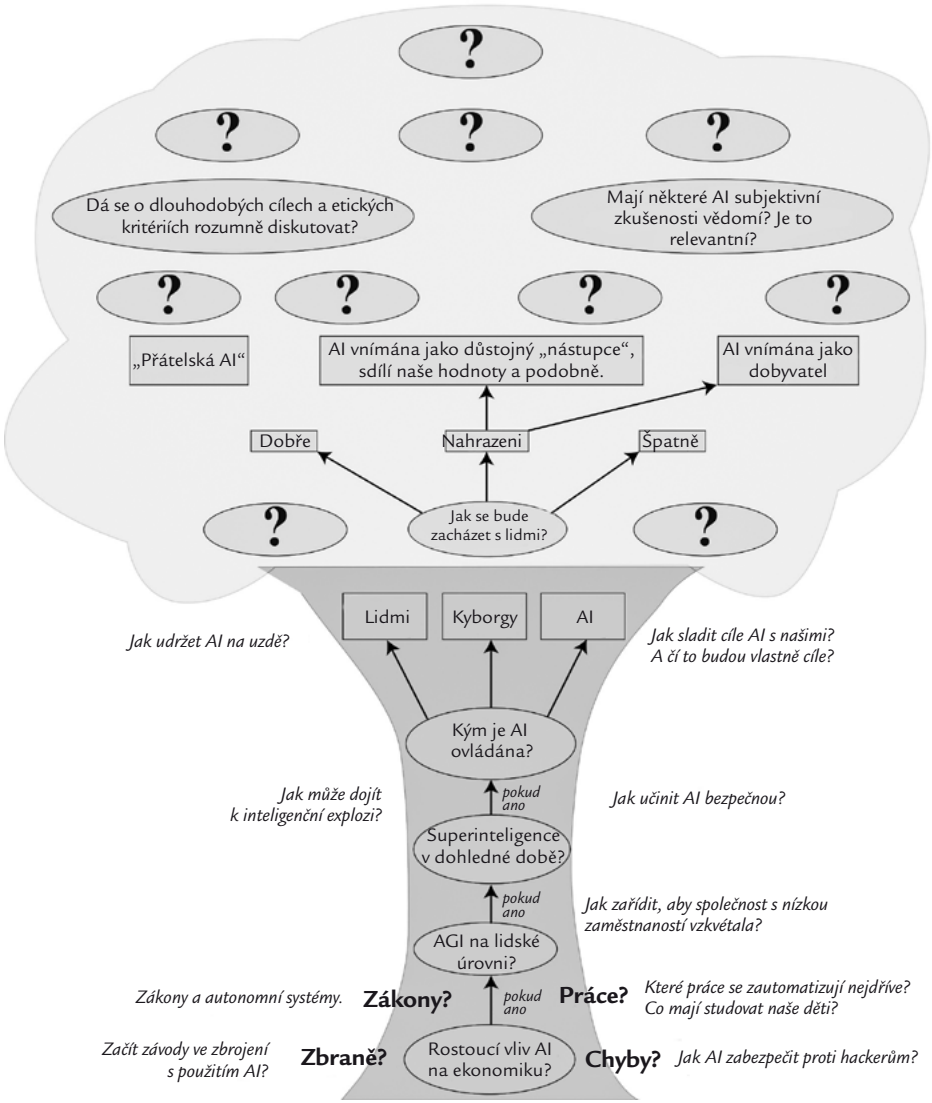
V kapitole 2 rozebereme základy inteligence a způsob, jímž se zdánlivě tupá hmota může přeskupit, aby si pamatovala, počítala a učila se. Jak budeme postupovat dál do budoucnosti, náš příběh se podle odpovědí na jisté klíčové otázky rozvětví do mnoha scénářů. Obrázek 1.6 shrnuje zásadní otázky, s nimiž se setkáme, když budeme kráčet dál časem k potenciálně stále pokročilejší AI.

Právě teď stojíme před volbou, zda začít závod ve zbrojení s použitím AI a otázkami, jak zbavit zítřejší systémy AI chyb a zajistit jejich robustnost. Bude-li ekonomický vliv AI nadále růst, musíme také rozhodnout, jak modernizovat naše zákony a co poradit našim dětem při výběru povolání, aby se vyhnuly pracím, které budou záhy automatizované. Takové krátkodobé otázky probereme v kapitole 3.

Když AI dosáhne lidské úrovně, budeme se také muset ptát sami sebe, jak zajistit její bezpečnost, a jestli můžeme nebo chceme vytvořit společnost volného času, která vzkvétá bez pracovních míst. To nás přivádí k otázce, jestli může dostat AGI daleko za hranice lidských možností spíše inteligenční exploze nebo pomalý a postupný růst. Širokou škálu takových scénářů prozkoumáme v kapitole 4 a poté v kapitole 5 rozebereme spektrum možností následků, které by to mělo, od zjevně dystopických po zcela utopické. Velí tu lidé, umělá inteligence, nebo kyborgové? Zachází se s lidmi dobře, nebo špatně? Nahradili nás? A pokud ano, vnímáme ty, kdo nás nahradili, jako dobyvatele, nebo jako důstojné nástupce? Moc mě zajímá, kterému ze scénářů v kapitole 5 dáváte přednost vy osobně! Založil jsem internetovou stránku, http://AgeOfAi.org, kde můžete sdílet své názory a zapojit se do diskuse.

Na závěr potom v kapitole 6 poskočíme miliardy let do budoucnosti, o níž paradoxně můžeme vyvozovat silnější závěry než v kapitolách předcházejících, protože konečné hranice života v našem kosmu nejsou stanoveny inteligencí, ale fyzikálními zákony.

Po završení naší pouti dějinami inteligence věnujeme zbytek knihy úvahám, o jakou budoucnost bychom měli usilovat a jak jí dosáhnout. Abychom dokázali spojit holá fakta s otázkami účelu a smyslu, podíváme se na fyzický základ cílů v kapitole 7 a na vědomí v kapitole 8. A konečně se v epilogu zaměříme na otázku, co právě teď můžeme dělat, abychom pomohli vytvořit budoucnost, jakou chceme.



Obrázek 1.6: Na úrovni a směru pokroku v AI záleží, které otázky se stanou zajímavými.

V případě, že jako čtenáři rádi přeskakujete, rádi uslyšíte, že většina kapitol je relativně samostatná – stačí vám vstřebat terminologii a definice z této první kapitoly a začátku kapitoly následující. Pokud jste odborník na umělou inteligenci, můžete klidně přeskočit celou kapitolu 2 kromě definic inteligence na jejím začátku. Jestli se s AI teprve začínáte seznamovat, pak vám kapitoly 2 a 3 poskytnou argumenty, proč nelze kapitoly 4–6 jednoduše zavrhnout jako nereálnou science fiction. Obrázek 1.7 shrnuje, kam na škálu fakta–spekulace jednotlivé kapitoly patří.

Čeká nás fascinující cesta. Tak začněme!

	Kapitola	Téma	Status
	Předehra: Příběh týmu Omega	Na zamyšlení	Vysoce spekulativní
Historie inteligence	1. Nejzásadnější debata	Klíčové myšlenky, terminologie	Nepříliš spekulativní
	2. Hmota se stává inteligentní	Základy inteligence	
	3. Blízká budoucnost	Blízká budoucnost	
	4. Inteligenční exploze?	Scénáře pro superinteligenci	Vysoce spekulativní
	5. Dohra	Dalších 10 000 let	
	6. Náš vesmírný odkaz	Další miliardy let	
Historie smyslu	7. Cíle	Dějiny chování zaměřeného na cíl	Nepříliš spekulativní
	8. Vědomí	Přirozené a umělé vědomí	Spekulativní
	Epilog: Příběh týmu FLI	Co bychom měli dělat?	Nepříliš spekulativní

Obrázek 1.7: Struktura celé knihy.

SHRNUTÍ ZÁKLADNÍCH FAKTŮ:

- Život, definovaný jako proces, který dokáže udržet svou komplexnost a replikovat se, se může vyvíjet a projít třemi stadii: stadiem biologickým (v1.0), kde se jeho hardware a software vyvíjejí evolucí, stadiem kulturním (v2.0), kdy dokáže svůj software přizpůsobit učením, a stadiem technologickým (v3.0), kdy dokáže navrhovat i svůj hardware, čímž se stává pánem svého osudu.
- Umělá inteligence nám umožní spustit Život 3.0 ještě v tomto století. Objevila se fascinující debata, o jakou budoucnost bychom měli usilovat a jak jí dosáhnout. V těchto sporech existují tři hlavní tábory: technoskeptici, digitální utopisté a hnutí za bezpečnou AI.
- Technoskeptici považují konstrukci nadlidské AGI za tak obtížný úkol, že k tomu nedojde ještě stovky let. Dělat si s ní (a se Životem 3.0) starosti už teď je proto podle nich hloupé.
- Digitální utopisté to naopak pokládají za pravděpodobné už v 21. století a Život 3.0 z celého srdce vítají, považují ho za přirozený a žádoucí další krok ve vesmírné evoluci.
- Hnutí za bezpečnou AI se také domnívá, že to nastane v 21. století, ale pozitivní výsledek nepovažuje za zaručený: je třeba ho zajistit těžkou prací, jíž je výzkum bezpečnosti AI.
- Kromě těchto legitimních diskusí, kde se neshodnou špičkoví světoví odborníci, existují i nudné pseudokontroverze zapříčiněné nedorozuměním. Tak například: nikdy neztrácejte čas dohadováním se o „životě“, „inteligenci“ nebo „vědomí“, dokud se nepřesvědčíte, že vy i váš partner používáte tato slova pro označení stejné věci! Tato kniha pracuje s definicemi z tabulky 1.1.
- Vyvarujte se i častých mylných představ z obrázku 1.5: „Superinteligence kolem roku 2100 je nevyhnutelná/nemožná.“ „Jen luddité si dělají starosti s umělou inteligencí.“

„Znepokojující je, že umělá inteligence začne páchat zlo, případně nabude vědomí, a to už za pár let." „Hlavním problémem jsou roboti." „Umělá inteligence nemůže ovládat lidi a nemůže mít cíle."

- V kapitolách 2–6 se podrobněji zaměříme na příběh inteligence od jeho skromných počátků před miliardami let k možným scénářům vesmírné budoucnosti za miliardy let od této chvíle. Nejprve se podíváme na výzvy blízké budoucnosti, jako jsou pracovní místa, zbraně s umělou inteligencí a hledání cesty k umělému bytí na lidské úrovni (AGI). Poté probereme možnosti fascinujícího spektra potenciálních budoucností s inteligentními stroji.
- V kapitolách 7–9 přejdeme od prostého popisu faktů ke zkoumání cílů, vědomí a smyslu a podíváme se, co můžeme udělat právě teď, abychom pomohli vytvořit budoucnost, jakou chceme.
- Tato diskuse o budoucnosti života s umělou inteligencí je asi vůbec nejdůležitější diskusí naší doby.