

PŘÍRUČKA EFEKTIVNÍHO ALTRUISMU

Jak pomáhat srdcem i rozumem



Efektivní altruismus Česko

2025

Obsah

Předmluva	5
1 - Co je to efektivní altruismus?	7
2 - Jak hledat zlato	23
3 - Mezní dopad: Jak nejlépe využít další zdroje	31
4 - Svět je hrozný. Svět se velmi zlepšil. Svět lze velmi zlepšit..	34
5 - Proč snižovat existenční rizika	40
6 - Prevence katastrofálních pandemií.	55
7 - Umělá inteligence mění náš svět – je na nás všech, aby to dopadlo dobře	66
8 - Prevence katastrofy spojené s umělou inteligencí.	76
9 - Uchvácení moci prostřednictvím umělé inteligence	106
Vaše cesta k efektivnímu altruismu: Jak můžete pomoci co nejlépe	109
Doporučené zdroje.	111
Zdroje kapitol.	113

Předmluva

Z jednoho pohledu žijeme v zlatém věku – za posledních 30 let vyšla z extrémní chudoby více než miliarda lidí, dětská úmrtnost klesla o dvě třetiny a díky internetu může kdokoli studovat na nejlepších univerzitách a číst nejnovější výzkum.

Z druhého pohledu stojíme na pokraji technologií, které to vše mohou překazit. Přichází silná umělá inteligence, nové biotechnologie umožňují vytvoření patogenů a máme tak čím dál více technologií schopných ukončit lidskou civilizaci a způsobit úplné vyhynutí lidstva.

Nikdy v historii lidstva jsme neměli takovou moc pomáhat – a současně takovou moc všechno zničit. Jak využít tu první a zabránit té druhé? Jak poznat, co skutečně pomůže světu?

Efektivní altruismus je výzkumný projekt, který se snaží na podobné otázky odpovědět. Zároveň je to světová komunita lidí, kteří se tyto poznatky snaží aplikovat – ve výběru kariéry, v darování na charitu, v politických rozhodnutích. Nemáme definitivní odpovědi, ale máme pár užitečných nástrojů a konceptů k tomu, jak přemýšlet systematictěji o tom, jak pomáhat.

Co najdete v této příručce:

Vybrali jsme překlady textů, které podle nás poskytují dobrý úvod do užitečných konceptů k přemýšlení o podobných otázkách.

Struktura příručky:

- **Nástroje:** Mentální modely pro efektivní pomáhání. Kapitoly 1–3.
- **Priority:** Jak vybírat mezi problémy světa, např. současné utrpení vs. budoucí rizika. Kapitoly 4–5.
- **Některé z nejslibnějších oblastí:** Úvod do rizik z AI a biosecurity. Kapitoly 6–9.

Online verze:

- Online verzi této příručky najdete na efektivni-altruismus.cz/prirucka



Sakra,
Susie!
Došel nám
KORIANDR!

Kapitola 1

Co je to efektivní altruismus?

Úvod

Efektivní altruismus je projekt, který si klade za cíl nalézat nejlepší způsoby pomoci druhým a uskutečňovat je.

Jde o výzkumný obor, jehož cílem je identifikovat nejpálčivější problémy světa a najít jejich nejlepší řešení, a zároveň i o prakticky zaměřenou komunitu, která se snaží tyto poznatky využít ke konání dobra.

Tento projekt má význam, protože ač mnoho z dobročinných projektů moc nefunguje, některé jsou nesmírně efektivní. Některé dobročinné organizace například při využití stejného množství zdrojů pomohou stokrát, nebo i tisíckrát většímu počtu lidí než jiné.

To znamená, že když pečlivě promyslíme, jak nejlépe pomoci, zmůžeme toho při řešení největších světových problémů mnohem víc.

S efektivním altruismem přišli vědci a vědkyně z Oxfordské univerzity, ale rozšířil se po celém světě a dnes jej aplikují desítky tisíc lidí ve více než sedmdesáti zemích.¹

Lidé inspirovaní tímto konceptem se zabývají nejrůznějšími projekty – od financování nadace, která rozdala 200 milionů moskytiér proti malárii, přes vědecký výzkum budoucnosti AI až po kampaně na podporu opatření, která by zabránila další pandemii.

Tyto lidi nespojuje jedno konkrétní řešení světových problémů, ale způsob uvažování. Snaží se nalézt neobyčejně dobré formy pomoci, aby stejné množství úsilí vedlo k dosažení neobvykle významných výsledků. Uvedeme několik příkladů toho, čeho zatím dosáhli, a hodnot, které sdílí:

Původně vyšlo jako *What is effective altruism?* Na <https://tinyurl.com/yc49m9ft>. Zde zkráceno.

Jaké jsou praktické příklady efektivního altruismu?

Prevence příští pandemie

Proč právě tohle téma?

Lidé, kteří se zabývají efektivním altruismem, se obvykle snaží vyhledávat rozsáhlé, snadno řešitelné a neprávem opomíjené problémy². Cílem je najít v současných iniciativách největší mezery a zjistit tak, kde nejvíce pomůže, když se přidá další člověk. Jednou z oblastí, které tato kritéria podle všeho splňují, je prevence pandemií.

Už v roce 2014 vědci a vědkyně věnující se efektivnímu altruismu uvedli, že vzhledem k tomu, kolikrát jsme se nějaké velké pandemii vyhnuli jen těsně, je pravděpodobné, že k ní během našeho života dojde³.

Ve srovnání s jinými světovými problémy se ovšem na přípravu na takovou událost věnovalo – a nadále věnuje – velmi málo prostředků. USA na prevenci pandemií například vynakládají asi 8 miliard dolarů ročně, zatímco na boj proti terorismu padlo za posledních deset let ročně přibližně 280 miliard dolarů⁴.



Odvracet teroristické útoky je rozhodně důležité. Rozsah tohoto problému se ale zdá menší. Když například vezmeme v úvahu pouze počet úmrtí, při teroristických útocích bylo za posledních 50 let zabito přibližně 500 tisíc lidí. Jen covid-19 si ovšem vyžádal přes 21 milionů životů⁵ – a HIV/AIDS 40 milionů⁶.

Nehledě na to, že budoucí pandemie by snadno mohla být mnohem horší než covid-19. Není vyloučeno, že se objeví nemoc nakažlivější než varianta omikron a zároveň stejně smrtelná jako pravé neštovice⁵.

Když efektivní altruisté odhalí nějaký zásadní přehlížený problém, komunita pak hledá řešení, která by mohla mít výrazný dopad, a přitom jsou ostatními, kdo se tématem zabývají, opomíjena. Tím se dostáváme k následujícím příkladům.



Příklady uskutečněných projektů

V roce 2016 se nadace Open Philanthropy, která z myšlenek efektivního altruismu vychází, stala největším zdrojem financování organizace Johns Hopkins Center for Health Security. Ta se jako jedna z mála zabývá výzkumem, jehož cílem je nalézt účinnější opatření při pandemiích, a hrála významnou roli při reakci na covid-19⁷.

Když propukl covid-19, členové a členky komunity efektivních altruistů také založili neziskovou organizaci 1DaySooner, která propaguje „human challenge“ testy. Při těchto testech vakcín se zdraví dobrovolníci nechají chorobou dobrovolně nakazit, aby bylo možné očkování testovat téměř ihned. Organizace 1DaySooner byla téměř jediná, kdo tento postup prosazoval. Přihlásilo se jí více než 30 tisíc dobrovolníků a dobrovolnic⁸ a sehrála významnou úlohu v zahájení prvního human challenge testu s covidem-19 na světě. Až budeme čelit příští pandemii, tento model bude možné využít znovu.

Členky a členové komunity efektivního altruismu rovněž přispěli k vytvoření programu Apollo Programme for Biodefense. Jde o návrh politiky na odvrácení příští pandemie počítající s vynaložením několika miliard dolarů.

Poskytování základního zdravotnického materiálu v chudých zemích

Proč právě tohle téma?

Často se říká, že dobročinnost začíná doma. Pro efektivní altruisty ale dobročinnost začíná tam, kde lze pomoci nejvíc. To často znamená zaměřit se na lidi, které současný systém opomíjí – a ti od nás mnohdy bývají nejdál.

Více než 700 milionů lidí žije za méně než 1,9 \$ na den⁹.



Částka, kterou má k dispozici Američan blížící se k hranici chudoby, je naproti tomu dvacetinásobná, a průměrný vysokoškolsky vzdělaný Američan utratí přibližně 107násobek¹⁰.

To je ze světového hlediska řadí mezi 1,3 % lidí s nejvyšším příjmem¹¹. (Tyto částky jsou již upraveny vzhledem k tomu, že v chudých zemích lze za stejné peníze získat víc.)

Globální nerovnost je obrovská. Přesunem zdrojů k těm nejchudším na světě proto lze dosáhnout opravdu velkého množství dobra. V bohatších zemích, jako je USA nebo Velká Británie, jsou vlády na záchranu života ochotné vydat více než milion dolarů¹². Rozhodně to stojí za to, ale v nejchudších zemích světa jsou náklady na záchranu života výrazně nižší.

Organizace GiveWell provádí důkladný výzkum s cílem zjistit, které zdravotní a rozvojové projekty jsou nejlépe podloženy důkazy a nákladově nejefektivnější. Zjistila, že mnohé snahy o pomoc nefungují¹³, zatímco jiné, jako například poskytování moskytiér impregnovaných přípravky na hubení hmyzu, mohou zachránit život dítěte s průměrnými náklady 5 500 \$. To je 180násobně méně¹⁴.

Tyto jednoduché aktivity v oblasti zdraví jsou tak levné a efektivní, že se na jejich smysluplnosti shodnou i ti nejvýznačnější kritici rozvojové pomoci¹⁵.



Příklady uskutečněných projektů

Výzkum organizace GiveWell využilo přes 110 tisíc jednotlivých dárců, kteří přispěli na doporučené dobročinné projekty více než miliardou dolarů. Podpořili organizace jako například Against Malaria Foundation, která rozdala více než 200 milionů moskytiér napuštěných přípravky na hubení hmyzu. Tyto aktivity zachránily podle odhadů celkem 159 tisíc životů¹⁶.

Nejchudším lidem na světě lze kromě dobročinnosti pomoci také prostřednictvím podnikání. Technologická společnost Wave, kterou členové komunity efektivních altruistů založili, například umožňuje posílat peníze do několika afrických zemí rychleji a několikanásobně levněji než dosavadní služby. Obzvláště užitečné je to pro migranty, kteří posílají peníze rodinám doma. Službu využívá více než 800 tisíc lidí třeba v Keni, Ugandě nebo Senegal. Jen v Senegal ušetřila služba uživatelům na poplatcích za převody měn stovky milionů dolarů – přibližně 1 % HDP země¹⁷.

Pomoc při vytváření výzkumného oboru zabývajícího se sladováním hodnot AI s těmi lidskými

Proč právě tohle téma?

Lidé zabývající se efektivním altruismem se často zaměřují na témata, která se zdají neintuitivní, nepochopitelná nebo přehnaná. Důvodem však je, že věnovat se tématům, která ostatní přehlížejí (jsou-li všechny ostatní faktory stejné), má větší účinek. A taková témata bývají (téměř výhradně) neobvyklá. Jedním z příkladů je problém sladování hodnot AI s těmi lidskými.

Rozvoj umělé inteligence postupuje rychle. Nejpokročilejší systémy AI dnes dokážou omezeně vést rozhovor, řešit matematické úlohy na vysokoškolské úrovni, vysvětlovat vtipy, generovat z textu velmi realistické obrázky a na základní úrovni programovat¹⁸.

Ještě před deseti lety nic z toho možné nebylo.

Konečným cílem předních výzkumných center v oblasti AI je vytvořit AI, která bude srovnatelná s lidmi nebo je v řešení veškerých úkolů i předčí. Předpovídat budoucnost technologií je nesmírně těžké, ale z různých argumentů a průzkumů mínění odborníků vyplývá, že k tomu v tomto století spíše dojde, než nedojde. Ze standardních ekonomických modelů pak vyplývá, že pokud AI dosáhne lidské úrovně dovedností, technologický pokrok se nejspíš prudce zrychlí.

Výsledkem by pravděpodobně byla zásadní transformace, rozsahem nejspíš srovnatelná s průmyslovou revolucí v 19. století nebo ještě významnější. Pokud bychom ji zvládli dobře, mohla by přinést hojnost a blahobyt všem. Kdybychom ji nezvládli, mohla by vést k obrovské koncentraci moci v rukou hrstky vyvolených.

V nejhorším případě bychom mohli nad samotnými systémy AI ztratit kontrolu. Bytosti se schopnostmi dalece přesahujícími ty naše bychom nedokázali ovládat, a tak bychom svou budoucnost neměli v rukách o nic více ji nyní mají třeba šimpanzi.

Tento problém by tudíž měl zásadní dopad nejen na současnou generaci, ale i na všechny příští. Obzvláště naléhavé je to tudíž z longtermistické perspektivy. Podle tohoto myšlenkového směru je jednou z klíčových morálních priorit současnosti zlepšovat dlouhodobou budoucnost.

Problém, jak zajistit, aby se systémy AI nadále řídily lidskými hodnotami, i když se schopnostmi dostanou na naši úroveň (nebo nás překonají), se označuje jako problém sladování hodnot AI s těmi lidskými a k jeho vyřešení bude zapotřebí pokrok v informatice.

Přestože tato záležitost může mít historický význam, zabývá se jí jen pár stovek výzkumníků a výzkumnic, zatímco zdokonalování systémů AI se jich věnují desítky tisíc.¹⁹



Příklady skutečných projektů

Jednou z priorit je zkrátka informovat o tomto tématu více lidí. V roce 2014 vyšla kniha *Superintelligence*, která zdůvodňovala zásadní význam slad'ování hodnot AI s těmi lidskými a stala se bestsellerem v žebříčku New York Times.

Další prioritou je vytvoření vědeckého oboru, který se na tento problém zaměří. Průkopník AI Stuart Russell a další lidé věnující se efektivnímu altruismu například založili na Kalifornské univerzitě v Berkeley Centrum pro AI v souladu s člověkem (The Center for Human-Compatible AI). Tento výzkumný ústav usiluje o vytvoření nového paradigmatu vývoje AI, v rámci něž je hlavním cílem podpora lidských hodnot.

Další lidé přispěli k založení týmů věnujících se slad'ování hodnot AI s lidskými v předních centrech zaměřených na vývoj AI, jako je DeepMind nebo OpenAI, a v pracích jako např. Concrete Problems in AI Safety navrhli výzkumné programy v oblasti slad'ování hodnot.

Zrušení průmyslových velkochovů

Proč právě tohle téma?

Lidé věnující se efektivnímu altruismu se snaží rozšířit svůj okruh zájmu, a to nejen na lidi žijící ve vzdálených zemích nebo budoucí generace, ale také na zvířata která nejsou lidmi.

Téměř 10 miliard zvířat v USA každý rok žije a umírá ve velkochovech²⁰. Tato zvířata se často po celý život nemohou fyzicky otočit nebo jsou kastrována bez anestezie.

Mnoho lidí se shodne na tom, že bychom zvířatům neměli působit zbytečné utrpení, tuto pozornost ale většinou věnujeme útlukům pro domácí mazlíčky. Velkochovy ale v USA projde přibližně 1400krát více zvířat než útulky²¹.



Přesto získají americké útulky ročně asi 5 miliard dolarů, zatímco na kampaně za ukončení průmyslových chovů se vynaloží jen 97 milionů dolarů²².



Příklady uskutečněných projektů

Mezi možné strategie patří přesvědčování lidí. Sdružení Open Wing Alliance, které získalo významné finance od sponzorů inspirovaných efektivním altruismem, vytvořilo kampaně s cílem přimět velké firmy, aby se zavázaly k tomu, že přestanou nakupovat vejce slepic v klecových chovech. Zatím se k tomu zavázalo přes 2 200 firem, v důsledku čehož bylo klecí ušetřeno 100 milionů ptáků.²³

Další strategií je výroba alternativních bílkovin. Pokud by byly levnější a chutnější než maso z velkochovů, poptávka po tomto mase by mohla zmizet a velkochovy by skončily. Toto odvětví se snaží povzbudit organizace Good Food Institute, která pomáhá zakládat společnosti jako např. Dao Foods v Číně nebo Good Catch v USA, podněcuje velké firmy (např. JBS, což je největší výrobce masa na světě) k tomu, aby do odvětví vstoupily, a zajišťuje státní podporu ve výši desítek milionů dolarů²⁴.

Organizace Open Philanthropy se stala jedním z prvních investorů do společnosti Impossible Foods, která vytvořila Impossible Burger – plně veganský burger, který se chutí velmi přibližuje masu a je teď součástí nabídky řetězce Burger King.

Zlepšování rozhodování

Proč právě tohle téma?

Lidé, kteří chtějí konat dobro, často rádi problémy řeší přímo, protože vidět hmatatelné výsledky své činnosti přináší větší motivaci. Záleží ale na tom, aby na světě bylo lépe, ne abyste toho dosáhli vlastníma rukama. Zastánci efektivního altruismu se tudíž často snaží pomáhat nepřímo, posilováním možností jiných.

Příkladem je zlepšování rozhodování. Kdyby se lidé v klíčových pozicích – například

politici a političky, vůdčí osobnosti soukromého a terciárního sektoru nebo ti, kdo ve financujících orgánech rozhodují o udělování příspěvků – rozhodovali obecně lépe, společnost by byla schopnější řešit celou řadu budoucích problémů, ať už budou jakékoli.

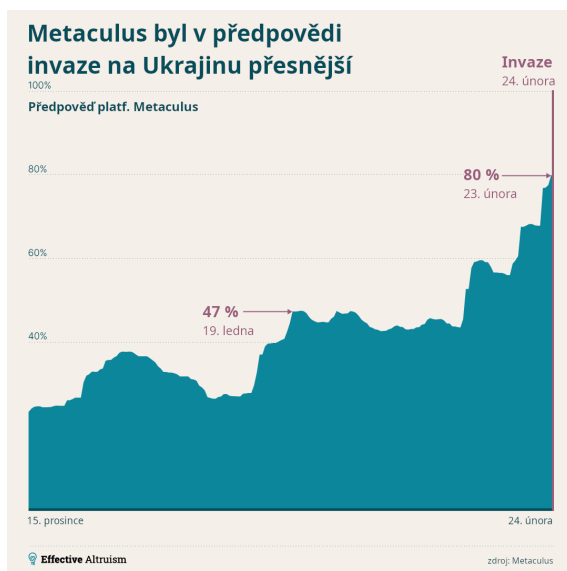
Pokud tedy nalezneme nové, opomíjené způsoby, jak zlepšit rozhodování důležitých aktérů, mohla by to být cesta k významným výsledkům. A zdá se, že existují slibné způsoby, jak toho dosáhnout.

Příklady skutečných projektů

Mnoho světových problémů se zhoršuje kvůli nedostatku důvěryhodných informací.

Technologická forecastingová platforma Metaculus se zaměřuje na důležité otázky (jako je například pravděpodobnost, že Rusko napadne Ukrajinu), agreguje předpovědi stovek forecasterů a foracasterek a udílí jim váhu podle přesnosti v minulosti. Pravděpodobnost ruské invaze na Ukrajinu byla v polovině ledna 2022 podle platformy Metaculus 47 % a krátce před invazí 24. února 80 %²⁴. Mnoho vědců, novinářů a expertů tehdy tvrdilo, že k ní určitě nedojde.

Výzkumný ústav Global Priorities Institute na Oxfordské univerzitě se zase věnuje základnímu výzkumu propojujícímu filozofii a ekonomii s cílem přicházet na to, jak mohou lidé činící nenápadná rozhodnutí rozpoznat nejpálčivější problémy. Ústav se podílel na vzniku nového vědního oboru spočívajícího ve výzkumu globálních priorit – vytvořil výzkumný program, zveřejnil desítky vědeckých prací a přispěl k podnícení příslušného výzkumu na univerzitách, mezi které patří Harvard, Newyorská univerzita, Texaská univerzita v Austinu, Yale, Princeton a další.



Jaké jsou zásady efektivního altruismu?

Výše uvedené projekty nejsou pro efektivní altruismus určující a oblast jeho zájmu se snadno může změnit. Určují ho naopak zásady, na kterých stojí jeho úsilí nalézt nejlepší způsoby pomoci druhým:

1. *Stanovování priorit*: Naše pocity při konání dobra obvykle opomíjí rozsah výsledků – když pomůžeme stovce lidí, jsme obvykle stejně spokojeni, jako když pomůžeme tisícovce. Protože ale některé formy dobročinnosti vedou k výrazně významnějším výsledkům než jiné, je zásadní snažit se pracovat s čísly a zhruba porovnat, kolik pomoci různé činnosti přinesou. Cílem není snažit se pomoci alespoň trochu, ale nalézt ty nejlepší cesty.
2. *Nestranný altruismus*: Je normální a rozumné, když člověku zvlášť záleží na vlastní rodině, přátelích nebo zemi. Když ale chceme vykonat co nejvíc dobra, snažíme se dávat stejnou váhu zájmům všech bez ohledu na to, kdy a kde žijí. Proto se soustředíme na ty nejopomíjenější skupiny. Do nich obvykle spadají ti, kteří na hájení vlastních zájmů nemají dostatek moci.
3. *Otevřeně hledání pravdy*: Důležitější než se v první řadě ztotožnit s nějakým tématem, společenstvím nebo přístupem, je zvážit mnoho různých způsobů pomoci a hledat ty nejlepší. To znamená věnovat mnoho času zvažování a promyšlení svých přesvědčení, být neustále zvědavý a otevřený novým důkazům a argumentům a dokázat případně názory výrazně změnit.
4. *Duch spolupráce*: Významnějších výsledků se často dá dosáhnout spoluprací. K tomu je třeba, aby byl člověk velmi čestný, zásadový a soucitný. Efektivní altruismus neznámá souhlasit s úvahou, že „účel světlí prostředky“. Znamená to být dobrým občanem a ambiciózně usilovat o lepší svět.

Tyto zásady nejsou nezpochybnitelné a mohou se měnit, jsou ale podle nás důležité a společností obecně nedocenené. Každý, kdo se při snaze hledat lepší způsoby pomoci druhým těmito zásadami řídí, se podílí na efektivním altruismu. Platí to bez ohledu na to, kolik času nebo peněz činnosti věnuje nebo na jaký problém se rozhodne zaměřit.

Efektivní altruismus lze přirovnat k vědecké metodě. Věda spočívá v tom, že při hledání pravdy se řídíme důkazy a rozumem – i když jsou výsledky neintuitivní nebo v rozporu s tradicí. Principem efektivního altruismu je, že důkazy a rozumem se řídíme při hledání nejlepších forem konání dobra.

Vědecká metoda je založena na prostých myšlenkách (např. že máte své názory testovat), vede ale k zásadně odlišnému pohledu na svět (např. kvantová mechanika). Stejně tak je na prostých myšlenkách založen i efektivní altruismus – že bychom se měli ke všem chovat stejně a že je lepší pomoci většímu než menšímu počtu lidí – ale vede k nekonvenčnímu a neustále se vyvíjejícímu pohledu na konání dobra.

Co můžete dělat vy?

Lidé, kteří se o efektivní altruismus zajímají, se nejčastěji snaží jeho myšlenky v životě uplatňovat následujícími způsoby:

- Volba profesí, které umožňují podílet se na řešení naléhavých problémů, nebo hledání způsobů, jak k tomu využít své stávající dovednosti. Lze se například řídit radami organizace 80,000 Hours.
- Přispívání pečlivě vybraným dobročinným organizacím, například s využitím výzkumu organizací GiveWell nebo Giving What We Can.
- Zakládání nových organizací, které přispívají k řešení naléhavých problémů.
- Podílení se na budování komunit, které se věnují palčivým problémům.

Delší seznam možností najdete na effectivealtruism.org/take-action.

Uvedený výčet není úplný. Efektivním altruismem se lze řídit, ať už se na konání dobra chcete soustředit do jakékoli míry a v kterékoli oblasti života. Ať chcete přispět jakkoli, jde o to, aby bylo vaše úsilí založeno na čtyřech hodnotách uvedených výše a abyste usilovali o co nejvyšší efektivitu svých činností.

Obvykle to zahrnuje snahu nalézat zásadní a opomíjené světové problémy, jejich nejefektivnější řešení a způsoby, jak byste k těmto řešením mohli přispět – bez ohledu na to, jaké množství času nebo peněz tomu chcete věnovat.

Díky takovému postupu a pečlivému rozvažování možná zjistíte, že s těmito zdroji lze dosáhnout mnohem větších výsledků. Skutečně je možné během své kariéry zachránit životy stovek lidí. Při spolupráci s dalšími členy komunity se pak můžete podílet na řešení některých z nejzásadnějších problémů, kterým dnes civilizace čelí.

Poznámky

1 Kde na světě se jednotlivé skupiny efektivních altruistů nachází, lze zjistit na fóru Effective Altruism Forum. Je tu uveden seznam skupin z více než 70 zemí.

2 Čím je problém méně přehlížený, tím spíše se těm nejlepším možnostem už někdo věnuje, a o to je těžší, aby činnost dalšího člověka měla účinek. Je pravděpodobné, že výnosy investice do nějakého problému jsou přibližně logaritmické. Z této logaritmické povahy výnosů vyplývá, že když se do nějaké oblasti investuje desetkrát víc než do jiné, další zdroje přinesou desetinový výsledek. V případě, že jde o dva stejně významné problémy, práce dalšího člověka, který se bude zabývat tím opomíjenějším, bude mít desetinasobný účinek.

3 openphilanthropy.org/research/biosecurity/

4 Odhaduje se, že na zdravotní bezpečnost vynaložily USA mezi roky 2010 a 2019 z rozpočtu 141 miliard \$. Máme za to, že na aktivity související s prevencí budoucích pandemií bylo použito 55 % těchto prostředků. 4 % například padla na řešení probíhající epidemie eboly, díky čemuž vznikla infrastruktura pro případ dalších pandemií. 17 % bylo nicméně vynaloženo na chemické hrozby a hrozby související s radioaktivním zářením tak, že na šíření budoucí pandemie to pravděpodobně nebude mít vliv.

$141 \text{ miliard} \times 0,55 = 79 \text{ miliard}$

Při přepočtu to za dané desetileté období vychází na 8 miliard dolarů ročně.

Federal funding for health security in FY2019 Watson, Crystal, et al., (2018): s. 281–303 (<https://www.liebertpub.com/doi/10.1089/hs.2018.0077>).

Organizace Open Philanthropy vede v patrnosti také další nadace a filantropie, kteří se tomuto tématu věnovali již před pandemií covidu-19. Máme za to, že dohromady vynaložili necelých 100 milionů dolarů.

Podle výpočtů ředitelky projektu Costs of War N. Crawford vydaly USA na boj proti terorismu mezi roky 2001 a 2022 5,8 bilionů dolarů.

5,8 bilionu : 20 let = 290 miliard \$ ročně.

United States budgetary costs of Post-9/11 wars Crawford, Neta C., Watson Institute for International & Public Affairs, Brown University, 2021 (<https://tinyurl.com/y2cgeduf>)

5 V období mezi lety 1970 a 2020 zemřelo v důsledku teroristických útoků přibližně 456 tisíc lidí. Zdrojem tohoto údaje je databáze Global Terrorism Database 2020.

Je třeba vzít v úvahu, že web Our World in Data uvádí: „Global Terrorism Database je nejkompaktnější dostupnější dataset týkající se teroristických útoků a údaje z nedávné doby tu jsou úplné. Na základě naší analýzy ovšem máme za to, že dlouhodobější údaje úplně nejsou (s výjimkou USA a Evropy). Nedoporučujeme proto vyvozovat z tohoto datasetu dlouhodobé trendy výskytu terorismu na světě.“

To znamená, že výše uvedený zdroj počet potvrzených úmrtí v důsledku terorismu pravděpodobně podhodnocuje. Ovšem i za předpokladu, že by počet úmrtí od roku 1970 byl přibližně stejný jako v desetiletí, kdy dosahoval nejvyšších hodnot (2010–2020), celkový počet obětí by stále byl pouze 1,2 milionu, tedy mnohem méně než počet obětí pandemií.

Úmrtí na covid-19:

Podle odhadu týdeníku The Economist dosahoval celkový počet nadúmrtí v důsledku covidu-19 v červnu 2022 21,47 milionu. Tento počet stále stoupá.

Tato data a model jsou dostupné na webu Our World in Data (<https://web.archive.org/web/20220728171227/https://ourworldindata.org/grapher/excess-deaths-cumulative-economist-single-entity>).

Máme za to, že jde o nejlepší současný odhad celkového počtu obětí covidu-19. Počet potvrzených úmrtí je nižší, přibližně 6 milionů, což ovšem nezahrnuje úmrtí nepřímá nebo nenahlášená. Metodika týdeníku Economist spočívá ve srovnání počtu nadúmrtí se sezónním průměrem, aby bylo možné odhadnout, kolik lidí zemřelo navíc, což se upraví o nenahlášená úmrtí.

Úmrtí v důsledku pandemií i terorismu jsou rozložena nerovnoměrně (s těžkým chvostem), takže minulé úmrtnosti obvykle vedou k podhodnocení rozsahu rizika.

Mohlo by například dojít k tomu, že teroristé odpálí jadernou zbraň ve velkém městě, což by mohlo zabít víc než milion lidí. V uplynulých padesáti letech k tomu nedošlo, kdyby se to ale stalo, byla by to hlavní příčina počtu obětí. Stejně tak mohlo v uplynulých padesáti letech dojít k mnohem horší pandemii, než byl covid-19 nebo HIV/AIDS.

Klíčovou otázkou pak je, jestli historické záznamy riziko více podhodnocují v případě terorismu, nebo v případě pandemie (tj. jestli má těžší chvost rozložení úmrtí v důsledku terorismu, nebo v důsledku pandemie).

Zdá se pravděpodobné, že nejhorší možný scénář je horší v případě pandemie než v případě terorismu. Není nijak vyloučeno, že nastane pandemie nakažlivější než covid-19, ale se smrtností 10–50 % nebo horší. A zdá se, že v historii k tomu už několikrát málem došlo.

Problém spočívající v tom, že ve vzorku chybí extrémní události, může být tudíž u pandemie horší než u terorismu. Koneckonců, nejpravděpodobnější způsob, jak by terorismus mohl zabít více než milion

lidí, nejspíš je způsobení pandemie.

Vzhledem k tomu, že na prevenci terorismu je vynakládáno asi stokrát víc prostředků než na prevenci pandemií, ačkoli pandemie v minulosti pravděpodobně způsobily desetkrát až stokrát více úmrtí, pro dosažení rovnoměrnější distribuce prostředků by muselo dojít k velmi výrazným úpravám ve prospěch pandemií.

Uvedená analýza zahrnovala pouze počty úmrtí, protože to je významná, ale zároveň relativně měřitelná metrika. Úmrtí způsobená pandemiemi a terorismem vedou také k významným nepřímým nákladům. Pro úplnější srovnání by bylo třeba pokusit se zahrnout jejich poměrný rozsah.

6 „Od začátku epidemie zemřelo na onemocnění související s AIDS 40,1 milionů (33,6 – 48,6 milionů) lidí.“ Global HIV & AIDS statistics — Fact sheet UNAIDS, 2022.

7 Open Philanthropy je nadace vycházející z myšlenek efektivního altruismu. Organizaci Johns Hopkins Centre for Health Security (CHS) poprvé podpořila v roce 2016. Následovalo několik dalších významných příspěvků včetně jednoho ve výši 16 milionů \$ v roce 2017 a dalšího ve výši 19,5 milionu \$ v roce 2019.

8 Ke dni 7. 7. 2022 šlo o 38 659 dobrovolníků a dobrovolnic.

9 Před covidem-19, v roce 2017, počet lidí žijících za méně než 1,9 \$ na den poklesl na 689 milionů. Z odhadů ale vyplývá, že dnes se poprvé od roku 1998 míra extrémní chudoby zvýšila. Odhaduje se, že méně než 1,9 \$ na den má teď 731 milionů lidí.

UN SDG 1 - End poverty in all its forms. Statistika OSN, 2022 (<https://unstats.un.org/sdgs/report/2021/goal-01/>)

Odhady byly přizpůsobeny tomu, že v chudých zemích si toho za tytéž peníze lze koupit víc (parita kupní síly). Ačkoli jsou tyto odhady v leccems problematické, je jasné, že příjem těsně nad úroveň existenčního minima mají stovky milionů lidí. Více podrobností případně najdete v článku *How accurately does anyone know the global distribution of income?* (<https://80000hours.org/2017/04/how-accurately-does-anyone-know-the-global-distribution-of-income/>).

10 V Česku má člověk na hranici chudoby 23x více než činí (starší) hranice chudoby 1,9 USD/den (tento počet je také přepočítán podle parity kupní síly - rozdílných cen v ČR oproti USA a rozvojovým zemím). Počet lidí v Česku pod touto hranicí je druhý nejmenší v OECD (po Dánsku). Vysokoškolsky vzdělaný člověk pak vydělává zhruba 70x více.

11 Hranice chudoby v USA na jednu osobu odpovídá ročnímu příjmu ve výši 13 590 \$.

$13\,590 : 365 = 37,23$ \$ na den.

To je dvacetinásobek mezinárodní hranice chudoby ve výši 1,9 \$ při přepočtení podle parity kupní síly.

Mediánový příjem lidí s vysokoškolským nebo vyšším vzděláním ve věku 25–26 pracujících na plný úvazek byl podle sčítání lidu v roce 2019 74 000 \$ ročně.

$74\,000 \$ / 365 = 202,7$ \$ na den.

$202 \$ / 1,9 = 107x$.

Jednočlenná domácnost v New Yorku s ročním příjmem 74 000 \$ hrubého má podle společnosti SmartAsset přibližně 53 000 \$ čistého.

Čistý příjem 53 000 \$ ročně vás podle kalkulačky organizace Giving What We Can řadí mezi 1,3 % lidí s nejvyšším příjmem na světě.

12 Britský Národní institut pro zdraví a klinickou kvalitou doporučuje v případě, že zákrok je spolehlivý, vynaložit na každý získaný rok života v plné kvalitě (QALY) až 30 000 £.

„V případě, že ICER (poměr inkrementálních nákladů a přínosů) na získaný QALY přesáhne 30 000 £, poradní orgány musí průkazněji doložit, že provedení zákroku představuje účelné vynaložení prostředků NHS (britského státního zdravotnického systému).“ Methods for the development of NICE public health guidance. UK National Institute for Health and Care Excellence, září 2012. (<https://tinyurl.com/24c4kqu7>).

V oblasti celosvětového zdraví se obvykle uvádí, že záchrana jednoho života odpovídá třiceti jednotkám QALY. Zdroj: Světová Banka (<https://tinyurl.com/2ydd3rom> Box 1.1)

Podle toho odpovídají náklady na záchranu života $30 \times 30\,000 \text{ £} = 900\,000 \text{ £} = 1,1 \text{ milionu \$}$.

V USA různé instituce odhadují „hodnotu života“ a tento ukazatel využívají k určování priorit jednotlivých projektů, na které je třeba vynaložit prostředky. Podle odhadu Federální agentury pro krizové řízení (FEMA) byla hodnota života v roce 2020 7,5 milionu \$. Tento odhad se v různých souvislostech mění. Podle odhadu Ministerstva dopravy USA byla například hodnota života v roce 2014 5,2 až 13,0 milionů \$.

13 <https://www.givewell.org/international/technical/criteria/impact/failure-stories>

14 Náklady na záchranu jednoho života odhadované organizací GiveWell se v čase mění (v závislosti na jejím výzkumu a dostupných možnostech), ale obvykle se pohybují mezi 2 500 a 7 500 \$. V roce 2021 bylo podle těchto odhadů možné rozdáváním moskytiér napuštěných látkou na hubení hmyzu zachránit jeden život s vynaložením 5 500 \$.

Nejaktuálnější odhady organizace najdete v její kompletní analýze nákladové efektivity: How We Produce Impact Estimates. GiveWell, červenec 2022.

15 <https://blog.givewell.org/2015/11/06/the-lack-of-controversy-over-well-targeted-aid/>

16 „Při cílení darů důvěřovalo doporučením GiveWell více než 110 tisíc dárců. Celkem poslali organizacím, které jsme doporučili, přes 1 miliardu \$. Tyto dary zachrání více než 150 tisíc životů a chudí lidé z celého světa tak získají peněžní příspěvky převyšující 175 milionů \$.“ Z webu About GiveWell, <https://www.givewell.org/about>

17 „Když jsme Wave spustili v Senegalu, náš průměrný převod by při využití největšího dosavadního poskytovatele mobilního bankovníctví byl třikrát až pětkrát dražší. Při vynásobení počtem lidí, kteří naši službu každý měsíc aktivně využívají, což jsou miliony, dojdeme k úsporám přes 200 milionů \$ ročně... a to je přibližně 1 % HDP Senegalu.“

Working at Wave is an extremely effective way to improve the world. Ben Kuhn, <https://www.wave.com/en/blog/world/>

18 Rozhovory: „Náš nejlepší model Meena vytrénovaný na principu end-to-end dosahuje... hodnoty SSA [Sensibleness and Specificity Average; průměrná smysluplnost a konkrétnost] 72 %... naše skóre SSA 72 % není daleko od SSA 86 % dosahovaného průměrným člověkem.“ Towards a Conversational Agent that Can Chat About...Anything. Adiwardana et al., Google, 28. 1. 2020.

Matematika: Minerva, model Googlu, podle grafů v odkazovaném článku správně řeší více než 50 % „středoškolských matematických soutěžních úloh“. Ostatní nejmodernější modely dosahovaly přesnosti pod 10 %.

Minerva: Solving Quantitative Reasoning Problems with Language Models. Dyer et. al, Google, 30. 6. 2022.

Vtípy: Model AI PaLM od Googlu je schopen vysvětlovat zcela nové vtípy, včetně těch, které nelze nalézt nikde na internetu.

Například:

„Vtip: Věděli jste, že Google zrovna do týmu pro TPU přijal mluvící velrybu? Ukázala jim, jak komunikují dva různé pody.“

Vysvětlení: TPU je typ počítačového čipu využívaného Googlem pro hluboké učení. „Pod“ označuje skupinu čipů TPU. „Pod“ je také skupina velryb (pozn.: anglicky „stádo“). Vtip spočívá v tom, že velryba dokáže komunikovat s dvěma skupinami velryb, ale mluví se tváří, jako že velryba je schopná komunikovat se dvěma skupinami TPU.“

Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance. Narang et al., Google, 4. 10. 2022. Archivovaný odkaz

Programování: Společnost Salesforce v kapitole 3.1 studie A Conversational Paradigm for Program Synthesis o svém AI nástroji CodeGen, který převádí lidské pokyny na kód, uvádí, že nástroj dosahuje v metrice HumanEval skóre 75 %, což znamená, že dokáže vyřešit 75 % úkolů v oblasti programování z datasetu HumanEval popsáných běžným lidským jazykem.

- 19** Odhadnout, kolik výzkumníc a výzkumníků se na nějaké téma zaměřuje, je náročné už proto, že se toto téma těžko definuje. Mnozí výzkumníci se zároveň zabývají více tématy a těžko se nastavuje práh pro to, co znamená „být výzkumníkem“. Tyto údaje je tudíž třeba brát jako odhady, které se mohou lišit až o přibližně trojnásobek a při různých interpretacích otázky i o řády.

V roce 2020 zveřejnilo v archivu vědeckých článků arXiv výzkumy zabývající se AI 87 tisíc autorů a autorek. Podle odhadu ve zprávě 2020 Global AI Talent Report společnosti Element AI se vývoji AI na celém světě věnuje ještě více lidí. Na sociálních sítích o sobě více než 155 tisíc osob uvádí, že pracují v oblasti výzkumu nebo vývoje AI. Předpokládáme ale, že někteří lidé pracující v oblasti vývoje AI se nezabývají dalším pokrokem této technologie. Vzali jsme tudíž v úvahu nejnižší odhad ve výši 87 000 a vydělili jsme ho přibližně dvěma. Výsledný odhad je tedy 40 tisíc.

Bezpečností AI se podle odhadu Gavina Leeche v roce 2021 zabývalo 270 až 830 lidí v ekvivalentech plného pracovního úvazku (<https://tinyurl.com/2bxktdff>). Horní hranice rozsahu tohoto odhadu ovšem vychází z pojetí výzkumu sladování hodnot AI s lidskými, které považujeme za příliš široké. Značná část tohoto množství je zároveň tvořena sčítáním času vynakládaného vědci, kteří se tomuto výzkumu věnují pouze okrajově. Naším cílem je ale určit, kolik výzkumníků a výzkumníc se na bezpečnost AI specializuje.

Iniciativa AI Watch se pokusila jednotlivé výzkumníky zabývající se bezpečností AI spočítat. Význačných vědců a vědkyň v tomto oboru našla 160. V tom je však zahrnuto i mnoho lidí, kteří o bezpečnosti AI více než rok nic nepublikovali, zatímco všichni ve výše uvedeném odhadu 87 tisíc nějaký výstup během uplynulého roku zveřejnili. Práh pro označení za „význačného“ vědce či vědkyni může být na druhou stranu vyšší než publikace v archivu arXiv.

Podle našeho výsledného odhadu se na bezpečnost AI zaměřuje 300 výzkumníc a výzkumníků.

- 20** V roce 2018 bylo v USA na maso poraženo 9,56 miliardy hospodářských zvířat. Tento počet se od té doby pravděpodobně zvýšil. Z toho 9,16 miliard bylo kuřat; 237 milionů krůt; 125 milionů prasat; 34 milionů hovězího dobytka a 2 miliony ovcí.
- 21** V roce 2021 prošlo útluky v USA přibližně 6,5 milionu zvířat. V roce 2011 to bylo 7,2 milionu. Za předpokladu trvalého úbytku to znamená, že v roce 2018 bylo v útlucích asi 6,7 milionů zvířat.
9,56 miliard : 6,7 miliony = 1427násobně víc zvířat v průmyslových chovech.
- 22** Financování zvířecích útluků:

Andrew Rowan spočítal, že v roce 2018 získaly tři tisíce nejvýznamnějších organizací provozujících útluky 5 miliard \$. Uvádí to v práci Cat Demographics & Impact on Wildlife in the USA, the UK,

Australia and New Zealand: Facts and Values Rowan et al. (2020), Journal of Applied Animal Ethics Research, s. 7–37.

Data, na kterých jsou tyto výpočty založené, nám Andrew Rowan v korespondenci potvrdil.

Financování kampaní za zrušení velkochovů:

Podle Open Philanthropy, získaly v roce 2018 organizace bojující za práva hospodářských zvířat následující částky:

Zavedené organizace v USA (PETA, PCRM, HSUS, ALDF, ASPCA): 32,3 milionů \$.

Nové významné organizace v USA (CIWF, WAP, RSPCA, HSI): 32,6 milionů \$.

Všechny další organizace v USA: 32,2 milionů \$.

$32,3 + 32,6 + 32,2 = 97,1$ milionů \$

- 23** Podle přehledu trhu vajec Ministerstva zemědělství USA je pouze v USA v květnu 2022 v bezklecových chovech 106,5 milionů slepic, zatímco v roce 2016 to bylo 17 milionů. Domníváme se, že díky práci sdružení Open Wing Alliance přešlo do bezklecových chovů dalších 100 milionů ptáků v Evropě, ovšem toto číslo nelze snadno připisovat konkrétně této skupině.
- 24** Vláda USA po dohodě s organizací Good Food Institute oznámila, že poskytne 10 milionů \$ na založení centra excelence pro buněčnou medicínu na Tuftsově univerzitě. Podle doporučení nezávislého auditu National Food Strategy ve Velké Británii by se mělo do výzkumu a inovací alternativních zdrojů bílkovin investovat 125 milionů £. Zdroj: GFI Year in Review 2021 (s. 3)

Kapitola 2

Jak hledat zlato

Owen Cotton-Barratt / 2016

V přednášce z roku 2016 popisuje Owen Cotton-Barratt z Oxfordské univerzity, jak mohou efektivní altruisté zlepšit svět, a používá k tomu příměr s hledáním zlata. Věnuje se řadě klíčových pojmů z oblasti efektivního altruismu, jako je rozdělení s těžkým chvostem, zákon klesajících výnosů nebo komparativní výhoda.

Tato verze Owenovy přednášky byla mírně upravena pro lepší srozumitelnost.

Efektivní altruismus jako kopání zlata

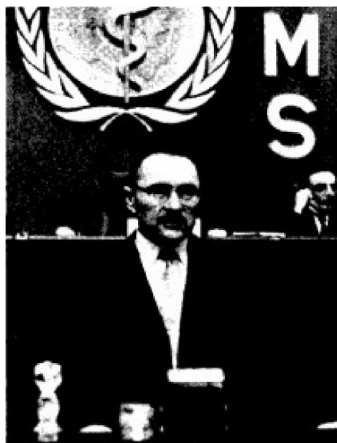


Obr. č. 1: Kopání zlata

Hlavní příměr celého článku spočívá v tom, že si efektivní altruismus představíme jako těžbu zlata. Budu na tomto přirovnání ilustrovat několik myšlenek. Zlato v tomhle případě představuje cokoli, na čem nám opravdu záleží. Můžete například usilovat o to, aby lidé byli šťastnější a vzdělanější, snažit se předcházet velkému utrpení nebo zvyšovat pravděpodobnost, že lidstvo vzlétne ke hvězdám. Při pohledu na slovo „zlato“

Původně vyšlo jako *Prospecting for gold* na tinyurl.com/3bff8pbv. Zde zkráceno.

se zkuste na chvilku zamyslet, na čem záleží vám (nemusí jít o jednu konkrétní věc), a místo zlata si představte právě to.



Obr. č. 2: Viktor Ždanov

Na obrázku č. 2 je fotografie Viktora Ždanova, o kterém jsem se dozvěděl z knihy Willa MacAskilla *Dobré úmysly nestačí*. Šlo o ukrajinského biologa, který měl zásadní podíl na vzniku programu vymýcení pravých neštovic. A tak mu nejspíš vděčíme za to, že nebyly ohroženy desítky milionů životů.

Něčeho takového pochopitelně nedosáhneme všichni. Na podobných příkladech ale vidíme, že někteří lidé vykopou mnohem víc zlata – tedy toho, čeho si altruisticky ceníme – než jiní. Je to dobrý důvod, abychom si kladli otázky jako například:

Jak to, že se některým lidem naskýtají lepší příležitosti než jiným? Co dělat, abychom takové příležitosti objevili také?

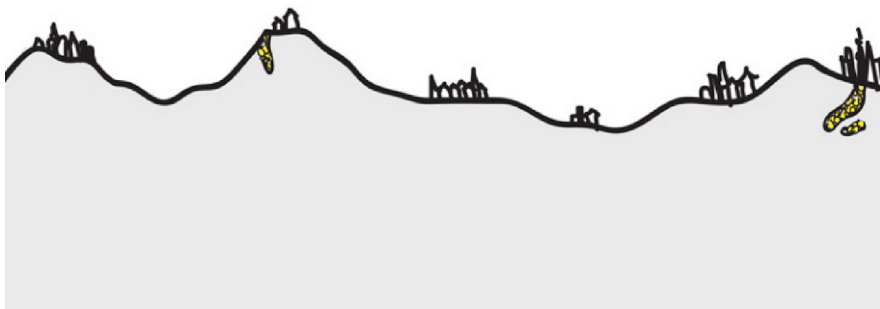


Obr. č. 3: Metody hledání zlata

Někteří lidé hovoří o tom, kde zlato je, a zabývají se mapami k jeho nalezení. Tomu já se v tomhle článku věnovat nebudu. Nebudu se vám snažit říct, kde se podle mě zlato nachází, ale zaměřím se na nástroje a techniky usnadňující jeho hledání.

Pro začátek bych rád vysvětlil, proč vůbec hovořím metaforicky. To, na čem nám záleží, je zásadní, složité a hodnotné, proč se to tedy snažím přirovnávat k pouhému zlatu? Důvodem je, že chci v tomhle článku popsat, jaké lze používat techniky, nástroje a postupy. Složitě hodnoty by od toho jen odváděly pozornost. Způsoby, jak hodnotné věci najít a získat je, se ale neliší, ať už si ceníme čehokoli. Myslím, že představit si místo nich něco jednoduchého pomáhá nenechat se svést k diskusi o tom, co vůbec hodnotné je.

Zlata není všude stejně

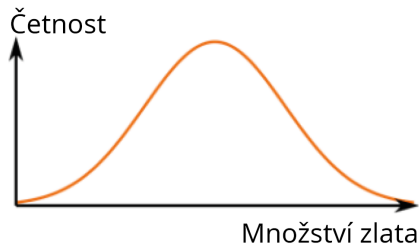


Obr. č. 4: Zlato je rozloženo nerovnoměrně

Jako první chci ilustrovat, že naše zlato – podobně jako to skutečné – je po světě rozprostřeno dost nerovnoměrně. Mnohde není skoro žádné a na pár místech je rozsáhlá zlatá žíla hluboko do země. Z toho vyplývá řada věcí. Jedna z nich je, že bychom velmi rádi našli právě tato ložiska.

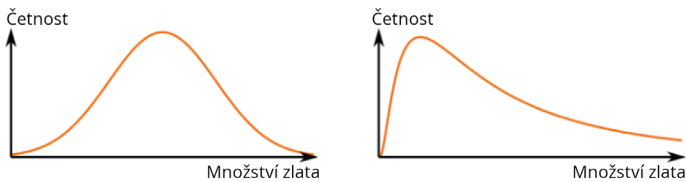
Rozdělení s těžkým chvostem

Další důsledek se týká získávání vzorků. Když mě například zajímá, jak bývají lidé vysocí, docela dobrý postup je jich pět změřit a spočítat jejich průměrnou výšku. Když chci ale zjistit, kolik je na světě průměrně zlata, zvolit pět náhodných míst a změřit množství zlata tam zas tak vhodné není. Je celkem pravděpodobné, že se trefím do pěti míst, kde není žádné, a tak jeho množství výrazně podhodnotím. Nebo se stane, že na jednom z nich bude zlata spousta, a já propadnu zavádějícímu dojmu, že ho je na světě hodně.



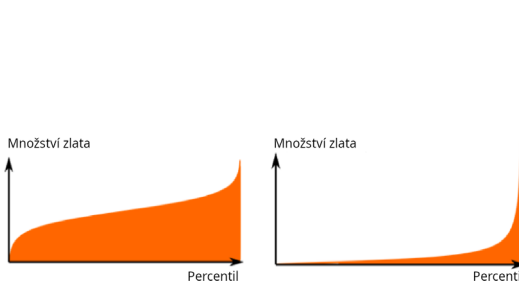
Obr. č. 5: Rozdělení bez těžkého chvostu

O této statistické vlastnosti nějakého jevu se často říká, že jeho rozdělení má těžký chvost. Na obrázku vlevo vidíme rozdělení, které těžký chvost nemá. Odpovídá situaci, kdy je na různých místech zlata různé množství, ale nikde ho není výrazně více nebo méně, než je běžné.



Obr. č. 6: Rozdělení bez těžkého chvostu a s ním

Na pravé straně je naopak rozdělení s těžkým chvostem. Vypadá podobně jako to nalevo, má však dlouhý chvost sahající k velkým hodnotám množství zlata, jejichž pravděpodobnosti neklesají příliš rychle. Z toho vyplývají různé důsledky.



Obr. č. 7: Rozdělení s těžkým chvostem

Jiný pohled na tato rozdělení nabízí obrázek č. 7. Tady jsem jednotlivá místa seřadil podle toho, kolik zlata se na nich nachází, vzestupně zleva doprava. Na vodorovné ose jsou percentily a na svislé množství zlata. Vybarvená plocha pod grafy v tomto případě odpovídá celkovému množství zlata. Vlevo, v případě rozdělení, které nemá těžký chvost, vidíme, že zlato je na různých místech rozloženo rovnoměrně. Kdybychom ho chtěli většinu získat, museli bychom kopat na co nejvíce místech.

Tak by se dala znázornit třeba solární elektrína. Na některá místa dopadá víc slunečního svitu než na jiná, ale kolik elektriny vygenerujete, závisí spíš na celkovém množství solárních panelů než na tom, kam přesně je umístíte.

Na grafu vpravo pak vidíme rozdělení s velkou plochou ve špičce na pravé straně. To znamená, že mnoho zlata – tedy toho, na čem nám záleží – je na nejvyšších percentilech rozdělení. Tam je ho mimořádné množství.

Nejsem geolog a o zlatě toho moc nevím, ale představuji si, že skutečné zlato je rozloženo jako na grafu s těžkým chvostem vpravo. *Můžeme si položit otázku, jestli něco podobného platí i pro příležitosti konat na světě dobro. Zde je pro to několik argumentů.*

Jevy s těžkým chvostem ve skutečném světě

Vznikají přirozeně
- např. logaritmicko-normální,
mocninné rozdělení



Obr. č. 8: Jevy s těžkým chvostem ve skutečném světě

Když se podíváme na svět v celé jeho složitosti, zjistíme, že případy rozdělení s těžkým chvostem jsou poměrně časté. V určitých situacích se dají některé typy rozdělení teoreticky očekávat. A empiricky zaznamenáme těžký chvost například v případě rozložení příjmů na světě. [Obr. č. 8]

Mnoho věcí tuto vlastnost pochopitelně nemá. S rostoucí složitostí systémů, kde probíhá mnoho interakcí, však stoupá i míra, v jaké se tento jev vyskytuje. A platí to i pro snahy zlepšovat svět.

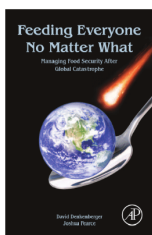
Různé důvody mě vedou k názoru, že se s tímto jevem setkáme i přímo u konkrétních

možností konání dobra.

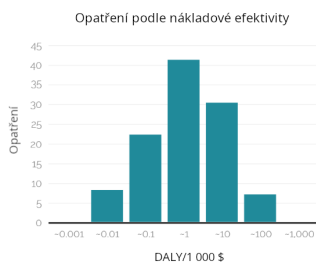
Jedním z těchto důvodů jsou zkrátka přesvědčivé argumenty. Pokud bych chtěl, aby lidé přestali trpět hladem – což si přeji – můžu si položit otázku: *Měl bych se věnovat přímo boji s hladomorem a snažit se poskytnout jídlo lidem, kteří dnes hladoví, nebo se orientovat víc do budoucnosti?* Podle některých názorů, které osobně považuji za přesvědčivé, může být efektivnější se věnovat výzkumu a bádání, jak zajistit potravu pro velké množství lidí pro případ, že se zhroutí zemědělství. Jde o extrémní příklad a obvykle takto neuvažujeme. Pokud ale chci zkrátka zajistit pro lidi potravu, některé způsoby se jeví výrazně efektivnější než jiné.

Těžký chvost u příležitosti ke konání dobra

Přesvědčivé argumenty



Data



Obr. č. 9: Jevy s těžkým chvostem u příležitosti ke konání dobra

Tato data z projektu Disease control Priorities 2 [Obr. č. 9] slouží k odhadu nákladové efektivity mnoha různých zdravotních opatření v rozvíjejícím se světě. Osa x využívá logaritmickou stupnici, takže opatření byla seskupena do kategorií a každý sloupec znázorňuje v průměru desetkrát vyšší efektivitu než ten na levé straně od něj. To znamená, že opatření ve sloupci úplně napravo jsou přibližně 10 000krát efektivnější než ta ve sloupci úplně nalevo. Jde o jednu konkrétní oblast světové zdravotní péče, kde se nám podařilo získat dostatek dat na to, abychom mohli takové odhady provádět. Vidíme, že rozsah možné nákladové efektivity je opravdu velký.

Vyplyvá z toho, že když chceme získat zlato, měli bychom se snažit hledat ložiska, kde ho je velké množství. To může vést k překvapivým zjištěním. Nemusí nás pak moc nadchnout, když zjistíme, že je něco na 90. percentilu. Dokud jsme o tom nic nevěděli, mohlo to být na kterémkoli místě grafu. Pokud se ale většina hodnoty nachází na 99. percentilu a my zjistíme, že něco je na 90., jde sice o užitečnou informaci, budeme však o dané věci mít horší mínění. To nastává v případech poměrně extrémního rozdělení a je zajímavé sledovat, jak k těmto neintuitivním jevům u rozdělení s těžkým chvostem může docházet.

Plyne z toho také, že kvůli problému se vzorkováním naivní empirismus („prostě uděláme spoustu věcí a uvidíme, co dopadne nejlépe“) nestačí. Abychom zjistili, jak efektivní něco skutečně je, není totiž možné získat dostatek vzorků a dostatečně změřit výsledky.

Když chceme co nejvíc zlata...

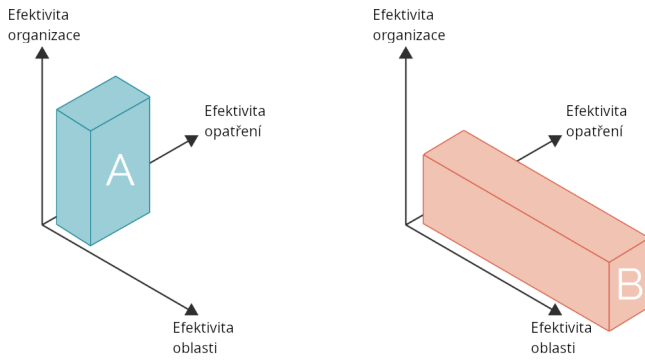


Obr. č. 10: Co nejvíc zlata...

Když chceme získat co nejvíc zlata, musíme jít tam, kde je ho hodně. Potřebujeme také vhodné nástroje k jeho těžbě a skvělé pracovníky, kteří tyto nástroje budou používat. Tuto analogii můžeme využít i pro příležitosti ke konání dobra. Efektivitu oblasti nebo aktivity, které se věnujeme, jde přibližně měřit. Efektivitu opatření, která v nějaké oblasti zavádíme, abychom vytvořili hodnotu, můžeme srovnat s jinými opatřeními v této oblasti. A zrovna tak lze efektivitu pracovníků nebo organizace, která opatření provádí, porovnávat s tím, jak by je provedli jiní pracovníci.

Hodnota je zhruba násobkem parametrů

Když tedy máme tyto různé parametry, je pak celková hodnota práce jejich násobkem. Na obrázku č. 11 je to znázorněno jako objem – a ten potřebujeme co největší. To znamená, že všechny parametry musí fungovat slušně, nebo alespoň žádný z nich nesmí být úplně špatný. Může z toho také plynout, že když se provádí skvělé opatření ve vhodné oblasti, ale pracuje na něm slabý tým, je někdy lepší ho nepodpořit a usilovat o to, aby tuto práci dělal někdo jiný. Nebo ten tým můžeme naopak něčím výrazně zkvalitnit. Podobně nemá smysl podporovat skvělý tým, pokud se věnuje něčemu, co se nezdá důležité.



Obr. č. 11: Hodnota je zhruba násobkem parametrů

Kapitola 3

Mezní dopad: Jak nejlépe využít další zdroje

Mezní dopad je účinek, který váš konkrétní vložený čas, finance nebo úsilí přinesou navíc. Místo abyste se soustředili na celkový dopad organizace nebo hnutí, zajímá vás, nakolik k tomu, na čem se už pracuje, přispěje vaše činnost. Můžete se tak lépe rozhodnout, kam své zdroje směřovat, aby přinesly účinek co největší.

Jaký je můj mezní dopad?

Lákavá často bývají velká hnutí s velkou hnací silou – být součástí něčeho, co mění svět, je přece strhující. Pokud ale chcete dosáhnout co největšího účinku, prostě se přidat k rozsáhlému hnutí s velkým dopadem nemusí být ta nejlepší cesta.

Mezní dopad vaší činnosti spočívá v jejím přírůstkovém účinku. Představte si například, že společnost vyrábějící toustovače se rozhoduje, jestli vyrobí ještě jeden další. Firmě se celkově může dařit, ale pokud je trh už toustovači nasycený, výroba dalšího by mohla vyústit ve ztrátu. Přestože celkové zisky společnosti jsou stále vysoké, mezní zisk (z příslušného dalšího toustovače) by v tomto případě byl negativní – takže jeho výroba nemá smysl.

Stejně tak je při vynakládání vlastního času, peněz nebo zdrojů na nějakou věc důležité se zaměřit na to, jaký to bude mít přidaný účinek. Snáze se tak vyhnete tzv. *efektu utopených nákladů*, kdy do něčeho nadále investujete jen proto, že to mělo úspěch v minulosti, i když budoucí příspěvky už takový účinek nemají.

Proč je to důležité?

Opravdu ale na téhle nuanci v životě záleží? Pokud chci například přispět dobročinně

Původně vyšlo jako *Marginal Impact: Making the Most of Additional Effort* na probablygood.org/core-concepts/marginal-impact/.

organizaci, není současný dopad její činnosti vhodným parametrem k určení účinku budoucích příspěvků? Ne nutně.

Milliony lidí se každoročně rozhodnou přispět Wikipedii. Dává to smysl – Každý měsíc ji využije více než miliarda lidí. Platforma jim poskytuje přístup k výukovým podkladům i odpovědi na praktické otázky a bojuje proti dezinformacím. Udržování její technické infrastruktury přitom stojí přibližně 36 milionů \$ ročně. Wikipedie tuto hodnotu tedy poskytuje s náklady nižšími než jeden cent měsíčně na člověka. Může snad mít nějaký příspěvek lepší dopad? Jenže i když Wikipedie ke svému fungování potřebuje uvedenou částku, poslední dobou získává každý rok příspěvky dalece přesahující 100 milionů \$.

Důležitější však je, že podle bývalé ředitelky¹ se většina těchto dodatečných příjmů investuje do projektů s vysokými náklady a nejasnými výsledky². Proto není jisté, zda příspěvky navíc pro organizaci Wikimedia vedou ke zlepšování obsahu Wikipedie. Je to příklad situace, kdy celkový dopad (nebo i celková nákladová efektivita) je mizerným parametrem na určení mezního dopadu dalších příspěvků. Prvních několik milionů dolarů, které Wikipedie dostane, je nesmírně cenných a důležitých, ale u těch jde už o hotovou věc – vy můžete rozhodovat jen o mezním dopadu stamilionového dolaru a dalších.

Tento efekt se netýká jen Wikipedie. Většina organizací podléhá zákonu klesajících výnosů (nebo přesněji klesajících mezních výnosů). To znamená, že první utracené dolary mají mnohem větší hodnotu než další investice. Mezní výnosy ale mohou být i nadprůměrné – například když má organizace značné fixní náklady nebo úspory z rozsahu. Jádrem věci tkví v tom, že při rozhodování o našich dalších krocích není vhodné se soustředit na celkový nebo průměrný dopad daných organizací – ať už je vyšší, nebo nižší – ale na mezní dopad našeho jednání.

Co to znamená pro výběr profesního směřování?

V případě volby kariéry je pojem mezního dopadu stejně důležitý jako při rozhodování, komu přispějeme finančně. *Mezní dopad* nového zaměstnance společnosti se může výrazně lišit od průměrného dopadu zbylého personálu a váš přínos závisí na různých faktorech. Rozhodování o vašem profesním směřování to může ovlivnit následovně:

Na první pohled se zdá, že nejlepší vždy je získat místo ve velmi úspěšné vlivné organizaci. Ve velké zavedené organizaci ale máte menší přímý vliv na její celkové směřování nebo úspěch. Takže váš mezní dopad nemusí být jasně patrný.

Přesto je ale práce v úspěšné organizaci často velmi dobrá volba. Tyto organizace obvykle vytvářejí skvělé příležitosti k práci jednotlivců, která má efekt, a poskytují cenný kariérní kapitál. Je ale důležité, aby celkový dopad firmy nezastínil vaše povědomí o tom,

1 <https://tinyurl.com/2yrejm8c>

2 <https://tinyurl.com/243tfdgn>

zda má efekt činnost na vaší konkrétní pozici.

Zároveň je důležité, abyste se na pozici hodili a dovedli se chopit jedinečných možností, které ostatní přehlíží. S obecně známými příležitostmi se sice často pojí větší celkový dopad, ale nalezení úzce specializované pozice, kde můžete svými konkrétními dovednostmi nebo nápady něco skutečně změnit, může vést k většímu meznímu dopadu.

Závěr

Když se při rozhodování, na co vynaložíte čas, peníze nebo profesní zájem, zaměříte na mezní dopad, pomůže vám to vyvarovat se nástrah orientace podle celkového nebo průměrného dopadu. Soustředění na to, jak můžete přispět navíc, vám umožní se vyhnout nadbytečným investicím v oblastech, kde budoucí potenciál není tak zářný jako minulé úspěchy, a nalézt takové příležitosti, aby vaše zapojení přineslo největší užitek.

Kapitola 4

Svět je hrozný. Svět se velmi zlepšil. Svět lze velmi zlepšit.

Max Roser / 2018, aktualizováno 2024

Je chyba si myslet, že tato tvrzení si odporují. Abychom přijali, že svět může být lepší, musíme si uvědomit, že platí všechna tři najednou.

Svět je hrozný. Svět se velmi zlepšil. Svět lze velmi zlepšit. Všechna tato tři tvrzení platí současně.

V diskusích o stavu světa se příliš často soustředíme na to první: Ve zprávách se klade důraz na to, co se nedaří, a málokdy se mluví o pozitivních změnách u nás nebo ve světě.

Vymezování se proti tomuto narativu přináší druhý extrém, který je zrovna tak škodlivý. Mluvit jen o pokroku ve světě, a přitom přehlížet obtíže, se kterými se lidé potýkají, nijak nepomáhá, nebo to přímo budí odpor.

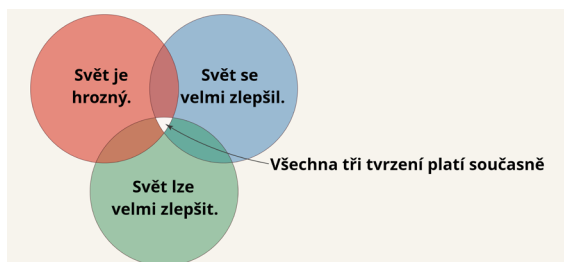
Když vidíme pouze problémy a slyšíme jen, co všechno je špatně, nemáme naději, že se to v budoucnu může zlepšit. Když ale slyšíme jen o pokroku a o tom, co se daří, stáváme se lhostejnými a ztrácíme ze zřetele problémy, kterým svět čelí. Oba tyto úzké pohledy mají tentýž důsledek: vedou nás k nečinnosti – paralyzují nás.

Nepoddát se ani jednomu z nich je náročné. Abychom si ale uvědomili, že svět lze zlepšit, potřebujeme chápat, že oba názory platí současně: svět je hrozný místo, a přitom se velmi zlepšil.

Pro ilustraci, co mám na mysli, uvedu jako příklad jednu z největších tragédií lidstva: každodenní úmrtí tisíců dětí.

Původně vyšlo jako *The world is awful. The world is much better. The world can be much better* na tinyurl.com/ynvmwre

Co platí pro dětskou úmrtnost, platí i pro mnoho dalších velkých problémů. Lidstvo čelí mnoha problémům, které se postupem času zmírnily, ačkoliv jsou stále hrozné, a zároveň víme, že se situace může dále zlepšovat.¹



Svět je hrozný

Podle nejaktuálnějších dostupných údajů z roku 2021 zemře 4,4 % všech dětí na světě před dosažením 15 let.

Znamená to, že každý rok zemře 5,9 milionů dětí. Každý průměrný den jich zemře 16 tisíc, tedy 11 každou minutu.

Svět, kde každý den dojde k tisícům tragédií, je pochopitelně hrozný.



1 V mnoha zásadních oblastech – pochopitelně ne ve všech – jsme dosáhli velmi významného pokroku. Příkladem je vzdělání, politická svoboda, násilí, chudoba, výživa a některé aspekty změny životního prostředí. Viz také můj krátký přehled historie životních podmínek ve světě (<https://tinyurl.com/yywds9nf>).

Svět se velmi zlepšil

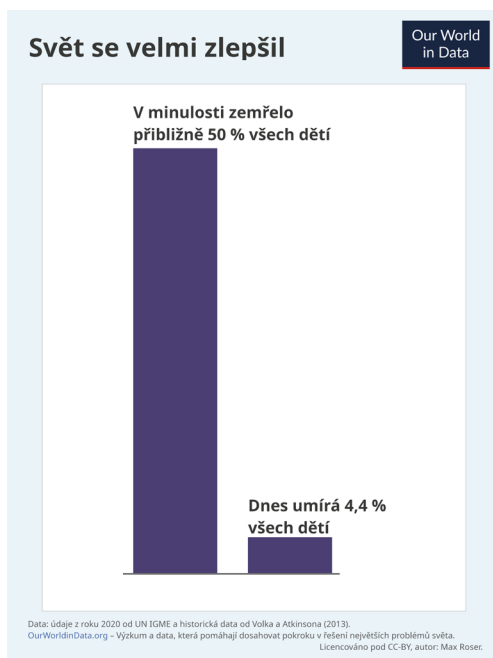
Dějiny přinášejí velké poučení, že věci se mění. Není ale snadné si představit, v jak strašných podmínkách se žilo v minulosti, takže je těžké pojmut, nakolik se svět změnil.

S povědomím o tomto měřítku nám mohou pomoci data. Podle odhadů historiků a historiček umírala v minulosti přibližně polovina všech dětí². Platilo to až do 19. století bez ohledu na to, kde na světě se dítě narodilo.

Je těžké si to představit, ale dětská úmrtnost na těch nejhorších místech dnes je mnohem nižší než v minulosti kdekoli. V Nígeru, zemi s aktuálně nejvyšší úmrtností, umírá přibližně 14 % všech dětí. Ještě před několika generacemi byla úmrtnost i na těch nejméně problematických místech třikrát vyšší.

Z dějin vidíme, že změnit svět lze. Dlouhodobá data o proměnách životních podmínek se ale bohužel ve škole probírají jen málokdy a v médiích se o nich příliš nemluví. Mnoho lidí proto bohužel ani o nejzákladnějších změnách světa k lepšímu vůbec neví.³

Tato skutečnost – že lze svět změnit a dosáhnout mimořádného pokroku pro celé společnosti – by však měla být známá všem. Pokud o nejsmyslnějších úspěších lidstva nevíme, není divu, že si příliš nevěříme a nevidíme naději na dosažení lepší budoucnosti.



2 <https://tinyurl.com/yene855f>

3 <https://tinyurl.com/yx8xzlff>

Svět lze velmi zlepšit

Postupný pokrok svědčí o tom, že v minulosti změnit svět šlo. Nabízí se ale otázka, zda to může pokračovat i do budoucna. Nenarodili jsme se právě v tom nešťastném momentu dějin, kdy se pokrok zastavil?

Při zkoumání dat o světě lze zjistit, že ne. Zlepšovat svět je pořád možné.

Presvědčit se o tom lze třeba při pohledu na ta místa na zemi, kde dnes panují nejlepší životní podmínky. Jsou důkazem, že velmi nízká dětská úmrtnost je nejen možná, ale už i reálná.

Regionem, kde mají děti nejvyšší šanci, že se dožijí dospělosti, je Evropská unie. Úmrtnost je zde 0,47 % – dospělosti se dožije 99,53 % všech dětí.

Chceme-li vědět, nakolik se svět může zlepšit, můžeme si položit otázku, jak by vypadal, kdyby to takhle bylo všude. Co kdyby se všem dětem na světě dařilo tak jako těm v EU? Znamenalo by to, že by jich každý rok zemřelo o pět milionů méně.

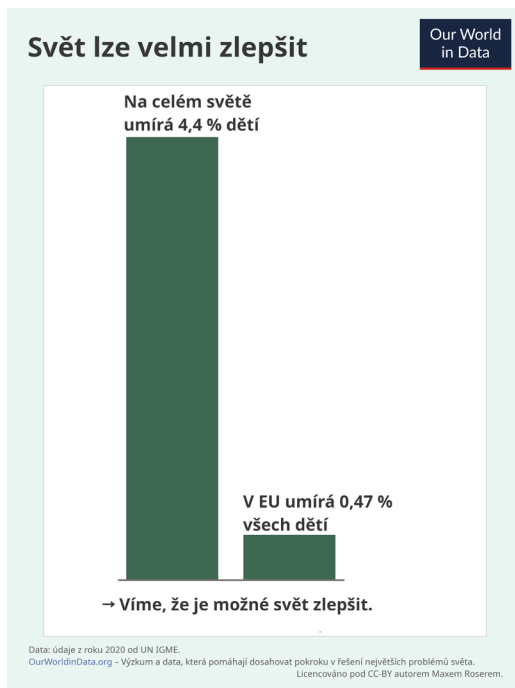
Celosvětový počet úmrtí by se snížil z 5,9 milionů na 0,6 milionu.

Dětská úmrtnost v EU je pochopitelně stále příliš vysoká a není důvod, aby se pokrok zastavil. Stovky dětí i v nejbohatších zemích dnes umírají na leukemii nebo nádory mozku. Musíme se tudíž snažit zjistit, jak těmto tragédiím předcházet.

Nejvíce možností, jak předcházet bolesti a utrpení dětí, je nicméně v chudších zemích. Víme totiž nejen, že situaci zde lze zlepšit, ale i jak toho dosáhnout.

Tento výzkum můžete ke zlepšování světa využít i vy sami. Doporučuji se řídit zjištěními neziskové organizace GiveWell.org. Její tým roky pracoval na rozpoznání těch nákladově nejefektivnějších dobročinných aktivit, aby vaše příspěvky měly na životy ostatních co největší pozitivní dopad. Některé z doporučených dobročinných projektů se věnují zlepšování dětského zdraví, takže máte příležitost přispět k pokroku ve snižování dětské úmrtnosti.

Předejít se dá úmrtím milionů dětí. Víme, že zlepšit svět je možné.



Svět je hrozný, a proto musíme vědět o pokrocích

Zprávy se často zaměřují na to, jak je svět hrozný. Děsit lidi je snazší než je nabádat, aby se snažili něco změnit k lepšímu, a špatným zprávám se vždy dostává spousty pozornosti.

Souhlasím, že je důležité vědět, co je na světě špatně. Vzhledem k rozsahu toho, co už jsme dokázali a co je možné v budoucnu, je však podle mě nezodpovědné informovat pouze o tom špatném.

Uvědomit si, že svět se zlepšuje, neznamená popírat, že čelíme velmi vážným problémům. Kdybychom už dosáhli nejlepšího možného světa, netrávil bych celé dny psaním a zkoumáním, jak jsme k tomu došli. Porozumět tomu, jak se svět zlepšil, je tak důležité proto, že je pořád strašný.

Doufám, že svou prací přispějí ke změně naší kultury, abychom brali možnost pokroku vážněji.

Jde o řešitelný problém: máme data a výzkum na to, abychom si uvědomili, jakým problémům čelíme a co s nimi lze dělat. Potíž je, že tato data a výzkum nevyužíváme. Data jsou často uložena v nepřístupných databázích a výsledky výzkumu zašifrované

do hantýrky vědeckých článků a mnohdy uzamčené za paypallem. Deset let buduji platformu Our World in Data, abych to změnil.

Jestli chceme, aby na zlepšování světa věnovalo energii a peníze více lidí, měli bychom mnohem více rozšířit informaci, že zlepšit svět je možné.

A proto musíme mít na paměti, že tři uvedená tvrzení platí současně.



Kapitola 5

Proč snižovat existenční rizika

Benjamin Todd / 2017, aktualizováno 2022

V roce 1939 napsal Einstein Rooseveltovi:

„Ve velkém množství uranu může být možné spustit jadernou řetězovou reakci... a je myslitelné – ačkoli mnohem méně jisté – že by tak bylo možné sestrojít nesmírně účinné bomby nového typu.“

O pouhých pár let později tyto bomby skutečně vznikly. Za méně než desetiletí jich bylo vyrobeno tolik, že hrstka aktérů mohla poprvé v dějinách zničit civilizaci.

Lidstvo vstoupilo do nové éry, kdy nečelíme pouze existenčním rizikům, vycházejícím z našeho přirozeného prostředí, ale také možnosti, že se zahubíme sami.

Co by v tomto novém věku mělo být pro naši civilizaci největší prioritou? Zdokonalovat technologie? Pomáhat chudým? Změnit politický systém?

Navrhujeme něco, o čem se nemluví zas tak často: nejdíc bychom měli usilovat o *přežití*.

Dokud bude civilizace existovat, budeme mít šanci vyřešit všechny ostatní problémy a dosáhnout mnohem lepší budoucnosti. Když ale vymřeme, bude to konec.

Proč se o této prioritě tolik nemluví? Jedním z důvodů je, že mnoho lidí změnu okolností stále nevnímá, a tak naši budoucnost za ohroženou nepovažuje.

Sociolog Spencer Greenberg udělal průzkum mezi Američany týkající se odhadu pravděpodobnosti, že lidstvo do padesáti let vymře. Zjistil, že podle mnohých je tato pravděpodobnost velmi nízká – více než 30 % tázaných ji odhadovalo na méně než 1 : 10 milionům.

Také jsme toto riziko považovali za velmi nízké, ale po jeho prozkoumání jsme názor změnil. Jak uvidíme, podle vědkyň a vědců, kteří se tímto tématem zabývají, je tato hrozba více než 1000násobně vyšší a nejspíš vzrůstá.

Tyto obavy vedly ke vzniku nového hnutí, které usiluje o ochranu civilizace. Připojili

se k němu Stephen Hawking, Max Tegmark a nové ústavy založené badateli a badatelkami na Cambridgeské univerzitě, MIT, Oxfordu a jinde.

Dále v tomto článku popíšeme největší rizika pro civilizaci včetně některých, která mohou být ještě závažnější než jaderná válka nebo změna klimatu. Potom se vás pokusíme přesvědčit, že snižovat tato rizika může být to nejdůležitější, čemu byste se mohli v životě věnovat. Pokud byste této problematice chtěli zasvětit své profesní směřování, můžeme vám také individuálně poradit.¹

Jaká je pravděpodobnost, že vás zabije asteroid? Přehled přírodních existenčních rizik

Pravděpodobnost 1 : 10 milionům, že v příštích 50 letech vyhyneme – což je předpoklad mnoha lidí – je nutně podhodnocená. Existenční rizika přirozeného původu lze poměrně přesně odhadnout z dějin a jsou mnohem vyšší.

Ke zničení civilizace by mohlo dojít, pokud by Zemi zasáhl asteroid o kilometrovém průměru. Astronomové pravděpodobnost takové srážky odhadují podle historických záznamů a sledování objektů ve vesmíru na přibližně 1 : 5000 na století. To je vyšší šance, než má běžný člověk, že zažije leteckou nehodu (asi 1 : 5 milionům na let), a již 1000násobně větší riziko než odhad 1 : 10 milionům udávaný některými lidmi.

Existují názory, že kilometrový objekt by sice způsobil katastrofu, ale ta by nebyla dostatečná na to, abychom vymřeli, takže uvedený odhad rizika je vysoký. Na druhou stranu ale existují i další přírodní hrozby, jako třeba supervulkány.

Přesto jsou přírodní rizika celkově malá. Podle chystaného článku Tobyho Orda nám úhrn všech přírodních rizik zvyšuje pravděpodobnost vyhynutí nejspíš o méně než 1 : 300 na století.

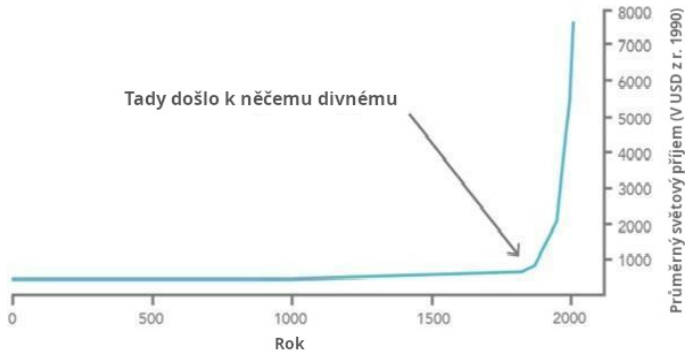
Jak ale brzy uvidíme, přírodní rizika bohužel blednou vedle těch způsobených lidmi. Právě proto je hrozba vyhynutí obzvlášť naléhavá.

Dějiny pokroku vedoucího k nejnebezpečnější éře v historii lidstva

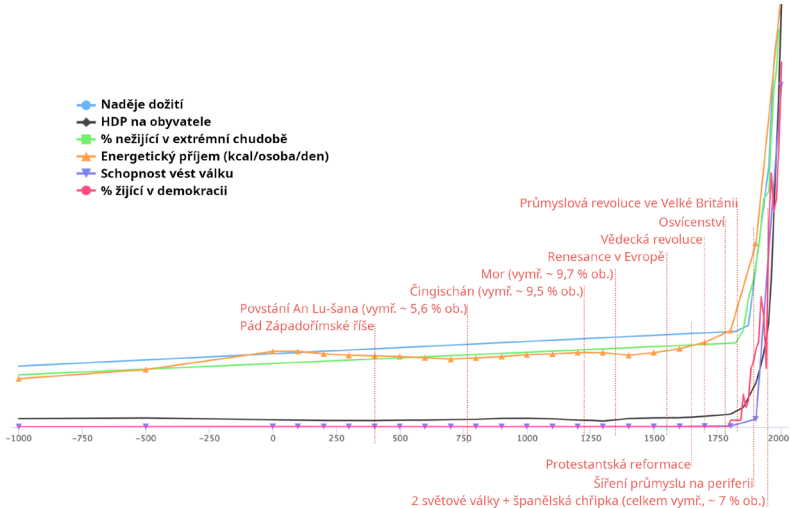
Když se podíváte na dějiny, zjistíte zejména, že po tisíciletí byli skoro všichni chudí a pak se to v 18. století změnilo.

Způsobila to průmyslová revoluce – nejspíš ta nejdůležitější událost vůbec.

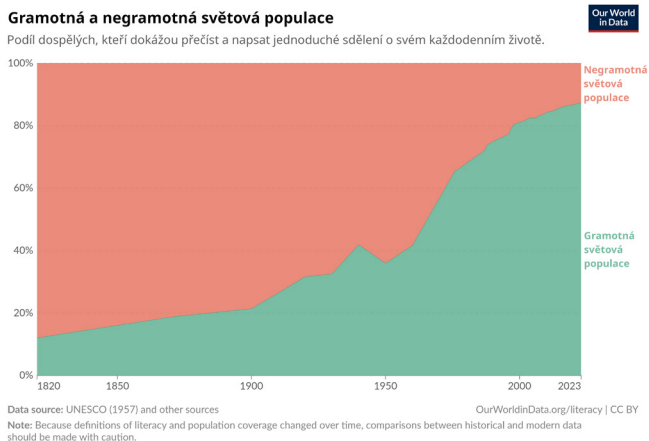
1 <https://80000hours.org/speak-with-us/>



Neroste jen bohatství. Na následujícím grafu je vidět, že v dlouhodobém měřítku rychle stoupá věk dožití, spotřeba elektriny a rozšiřuje se demokracie. Procento lidí žijících v chudobě naopak významně klesá.



Značně vzrůstá také gramotnost a vzdělanost:



Zdá se také, že spolu s bohatstvím roste i spokojenost.

Steven Pinker v knize *The Better Angels of Our Nature* uvádí, že klesá násilí. (V nedávném podcastu s Bearem Braumoellerem se nicméně zabýváme několika důvody, proč to tak možná není.²⁾

Narůstá svoboda jednotlivců a klesá rasismus, sexismus a homofobie.

Mnozí si myslí, že se světem je to čím dál horší, a je pravda, že moderní civilizace má na svědomí hrozné věci, jako třeba velkochovy. Jak ale ukazují data, mnohé významné ukazatele pokroku se zásadně zlepšily.

Přesněji řečeno, ať se v minulosti podle vás dělo cokoli, zdokonalování technologií a politické organizace a růst svobody vytváří při pohledu do budoucnosti pro naše potomky potenciál vyřešit současné problémy a mít se výrazně lépe. Je možné se vymáknit z chudoby, předejít klimatické změně, zmírnit utrpení a tak dále.

Na druhém grafu si ale také všimněte fialové čáry: schopnost vést válku. Je založená na odhadech celosvětové vojenské síly podle historika Iana Morrise – a také výrazně roste.

Problém je, že technologický rozvoj má obrovský přínos, ale vede také k obrovským rizikům.

Kdykoli objevíme novou technologii, většinou to přinese velké výhody. Existuje ale také možnost, že vytvoříme technologii s větší ničivou silou, než dokážeme rozumně využívat.

² <https://tinyurl.com/2d68r2kq>

A tak přestože současná generace prožívá nejblahobytnější období lidských dějin, žije možná zároveň v tom nejnebezpečnějším.

První ničivou technologií tohoto typu byly jaderné zbraně.

Jaderné zbraně: dějiny úniků o vlásek

Dnes nás všechny napadne jaderný program Severní Korey, ale současné události jsou jen jednou kapitolou v dlouhém příběhu, kdy mnohokrát chybělo málo.

Jen během karibské krize jsme byli krok od jaderné války hned několikrát. Jednou se Američané rozhodli, že pokud jim protivník sestřelí průzkumný letoun, zaútočí na Kubu rovnou, bez dalšího zasedání válečné rady. Následujícího dne byl jeden letoun sestřelen. J. F. Kennedy radu přesto svolal a ta se rozhodla nezaútočit.

Invaze by možná spustila jadernou válku – později se ukázalo, že Castro se přikláněl k jaderné odvetě, i kdyby „vedla k naprostému zničení Kuby.“ Někteří kubánští velitelé raket také měli samostatnou pravomoc v případě invaze zaútočit na americké vojsko taktickými jadernými zbraněmi.

Při jiném incidentu se ruská ponorka snažila na Kubu propašovat materiál, ale byla odhalena americkým námořním svazem. Lodě začaly shazovat slepé hlubinné pumy, aby ji donutily k vynoření. Ruský kapitán si myslel, že jde o nálože skutečné, a protože mu nefungovalo rádiové spojení, usoudil, že třetí světová válka už začala. Nařídil zaútočit na americká plavidla jaderným torpédem.

Naštěstí k tomu potřeboval schválení dalších vysokých důstojníků. Jeden z nich, Vasilij Archipov, nesouhlasil, a válku tak odvrátil. Díky, Vasiliji Archipove.



Vasilij Archipov

Kennedy později odhadoval, že při uvážení všech těchto událostí byla pravděpodobnost jaderné války celkově „někde mezi 1 : 3 a 1 : 1“.

Jak je patrné ze seznamu na této pěkné stránce na Wikipedii³, v Rusku to bylo nahnuté ještě mnohokrát, a to i po skončení studené války. A to jsou jen ty známé případy.

Dnes mají odborníci obavy nejen ze Severní Korey, ale také z napětí mezi Indií a Pákistánem, přičemž jaderné zbraně mají obě země.

Hlavní problém spočívá v tom, že několik zemí udržuje velký jaderný arzenál, který lze použít v řádu minut. Falešný poplach nebo nehoda může tudíž rychle eskalovat v otevřený konflikt, zvláště když jsou mezinárodní vztahy napjaté.

Způsobila by taková válka kolaps civilizace? Původně se mělo za to, že jaderný výbuch by mohl mít takovou teplotu, že by se vznítila atmosféra a Země by se stala neobyvatelnou. Podle odhadu vědců to bylo natolik nepravděpodobné, že testování zbraní považovali za „bezpečné“, a dnes víme, že k tomu dojít nemůže.

V 80. letech panovaly obavy, že popel z hořících budov by přivodil dlouhodobou zimu, která by na desítky let na Zemi znemožnila pěstování plodin. Podle moderních klimatických modelů je pravděpodobnost nukleární zimy tak tuhá, že by zahubila všechny, velmi nízká. Vzhledem k nejistotě v modelování to ale nelze tvrdit s jistotou.

I „mírná“ nukleární zima by ale mohla způsobit hladomor. Proto a z dalších důvodů by jaderná válka byla velmi destabilizující a není jisté, že by se po ní civilizace dala do pořádku.

Jak pravděpodobné je, že by taková válka civilizaci mohla navždy ukončit? Velmi těžko se to odhaduje, ale nelze říct, že by pravděpodobnost takové války byla v následujícím století méně než 0,3 %. To by znamenalo, že riziko představované jadernými zbraněmi je větší než všechna přírodní rizika dohromady. O jaderných rizicích více na ⁴.

Proto byla 50. léta pro lidstvo začátkem nové éry. Poprvé v dějinách začalo být možné, aby hrstka vlivných lidí zpusťovala celý svět. Největší hrozbou pro naše přežití jsme dnes my sami – proto žijeme v nejnebezpečnější době lidské historie.

Jaderné zbraně přitom nejsou jediným prostředkem, jak bychom mohli kolaps civilizace způsobit.

Jak velké riziko představuje změna klimatu?

Americký prezident Obama ve své Zprávě o stavu Unie v roce 2015 prohlásil, že „změna klimatu je největší hrozbou pro budoucí generace.“

Klimatická změna rozhodně představuje pro civilizaci zásadní riziko.

Nejpravděpodobnějším výsledkem je oteplení o 2 až 4 stupně, což by bylo špatné, ale náš druh by to mohl přežít.

Podle některých odhadů ale existuje 10% pravděpodobnost oteplení přes 6 °C, a možná i 1% pravděpodobnost oteplení o 9 °C.

Zdá se tedy, že rozsáhlá klimatická katastrofa způsobená CO₂ je podobně

3 https://en.wikipedia.org/wiki/List_of_nuclear_close_calls

4 <https://tinyurl.com/2loebk8k>

pravděpodobná jako jaderná válka.

Jak ale popisujeme ve článku věnovaném klimatické změně⁵, zdá se nepravděpodobné, že oteplení i o 13 °C by přímo způsobilo vymření lidstva. Badatelé zabývající se těmito problémy mají tudíž za to, že jaderná válka by nás přímo vyhubila pravděpodobněji, protože by mohla způsobit nukleární zimu – proto se domníváme, že jaderné zbraně představují ještě větší hrozbu než klimatická změna.

Přesto je ale změna klimatu zásadní problém a její destabilizující důsledky by mohly zhoršit další hrozby (včetně rizika jaderného konfliktu). To by mělo náš odhad tohoto rizika ještě zvýšit.

Které nové technologie můžou být stejně nebezpečné jako jaderné zbraně?

Vynález jaderných zbraní dal jen o pár desetiletí později, v 60. letech, vzniknout protijadernému hnutí a proti klimatické změně začali brzy bojovat environmentalisté.

Méně pozornosti se ale věnuje tomu, že nové technologie přinesou další hrozby. Proto potřebujeme hnutí, které se soustředí na ochranu civilizace obecně.

Předvídat budoucnost technologií není snadné, ale protože civilizaci máme jen jednu, je třeba dělat, co můžeme. Mezi kandidáty na příští technologii stejně nebezpečnou jako jaderné zbraně patří třeba tyto:

Uměle vytvořené pandemie

Mezi lety 1918 a 1919 zemřela více než 3 % světové populace na španělskou chřipku. Kdyby taková pandemie přišla dnes, kvůli rychlé dopravě po celém světě by se možná potlačovala ještě hůř.

Větší obavy ale budí, že brzy může být možné geneticky vyvinout virus nakažlivý jako španělská chřipka, ale ještě více smrtící, který by se zároveň mohl roky nepozorovaně šířit.

Šlo by o zbraň stejně ničivou jako ty jaderné, ale jejímu použití by se hůře předcházelo. K výrobě jaderných zbraní jsou zapotřebí velké továrny a obtížně dostupné suroviny, takže není tak složité je mít pod kontrolou. Umělé viry by ale mohlo vytvořit v laboratoři pár lidí s doktorátem z biologie. Deníku Guardian se v roce 2006 podařilo nechat si poslat segmenty vymýceného viru pravých neštovic poštou. O využití takových zbraní s nerozlišujícími účinky projevil zájem některé teroristické skupiny. Více o rizicích pandemie se dočtete v kapitole 6.

5 <https://tinyurl.com/yee83759>



Kdo si objednal neštovice?

Umělá inteligence

Další novou technologií s potenciálem získat obrovskou moc je umělá inteligence.

Jediným důvodem, proč svět neřídí šimpanzi, nýbrž lidé, je inteligence. Naše velké účinné mozky nám umožňují značnou vládu nad světem, přestože jsme fyzicky mnohem slabší než šimpanzi.

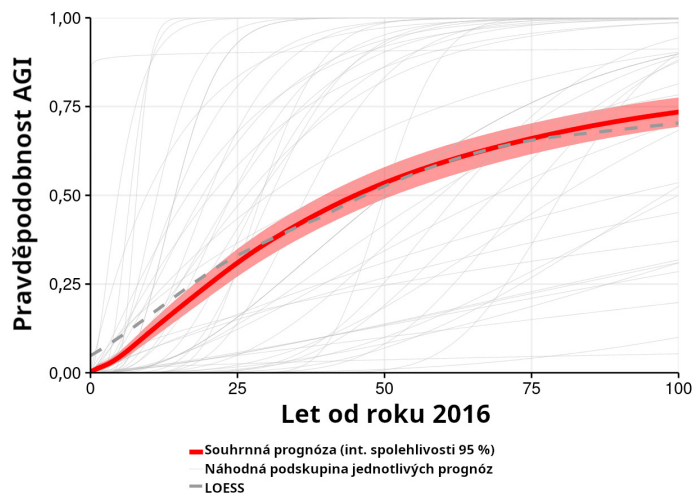
Co by se tudíž stalo, kdybychom jednoho dne vytvořili něco mnohem inteligentnějšího, než jsme my sami?

V roce 2017 proběhl na vrcholných konferencích průzkum mezi badateli a badatelkami, kterým vyšly recenzované vědecké články o umělé inteligenci. Byli dotázáni, kdy podle nich vyvineme počítače nadané inteligencí lidské úrovně – tedy stroje schopné vykonávat veškeré pracovní činnosti lépe než lidé.

Podle mediánového odhadu existuje 50% pravděpodobnost, že AI lidské úrovně vyvineme za 45 let, a 75% pravděpodobnost, že to bude do konce století.

Tyto pravděpodobnosti se odhadují těžko a vědci a vědkyně udávali velmi odlišné hodnoty podle toho, jak přesně otázka zněla. Zdá se nicméně, že existuje přinejmenším rozumná pravděpodobnost, že dojde k vynálezu nějakého typu transformativní strojové inteligence v příštím století. Větší nejistota navíc značí, že může být vynalezena spíš dříve, než se očekává, než později.

Jaká rizika může tento vývoj přinést? Obavy z hrozeb představovaných mocnými počítačovými systémy vyjádřili už raní průkopníci informatiky jako Alan Turing a Marvin Minsky a tyto hrozby stále existují. Řeč není o tom, že by počítače „propadly zlu.“ Obavy vzbuzuje spíše to, že by jedna skupina mohla mocný systém AI využít k ovládnutí světa nebo by došlo k nějakému jinému zneužití AI. Kdyby Sovětský svaz vyvinul jaderné zbraně o deset let dříve než USA, mohl by získat vládu nad celým světem. Mocná počítačová technologie by mohla představovat podobná rizika.



Další problém je, že spuštění takového systému by mohlo mít nečekané důsledky, protože předvídat chování něčeho chytřejšího, než jsme my, by bylo obtížné. Dostatečně mocný systém by také mohlo být těžké držet pod kontrolou, a tudíž ho po uvedení do chodu upravit. Těmto otázkám se věnuje oxfordský profesor Nick Bostrom ve své knize *Superintelligence* a také průkopník AI Stuart Russell.

Většina odborníků a odbornic je toho názoru, že dokonalejší AI přinese kladnou změnu, ale shodují se, že to má rizika. Ve výše uvedeném průzkumu respondenti odhadovali 10% pravděpodobnost, že dopad vývoje AI lidské úrovni bude „špatný“ a 5% pravděpodobnost, že bude „extrémně špatný“, např. vyhynutí lidstva. Přitom bychom měli nejspíš předpokládat, že u této skupiny bude docházet k pozitivnímu zkreslení, protože se AI živí.

Z kombinace těchto odhadů vyplývá, že pokud je pravděpodobnost vzniku vysokoúrovňové strojové inteligence v příštích sto letech 75 %, představuje pravděpodobnost velkého neštěstí způsobeného AI 5 % ze 75 %, tudíž přibližně 4 %. (Více o umělé inteligenci v kapitolách 7–9.)

Další rizika budoucích technologií

Někteří vyjadřují obavy i z dalších nových technologií, například některých forem geoinženýrství a nanotechnologií. Zdá se ale, že ty nejsou tak na dosah jako technologie popsané výše, a tak je lidé nepovažují za tak nebezpečné. Delší seznam existenčních rizik najdete na nickbostrom.com/existential/risks.

Znepokojivější jsou ale nejspíš rizika, na která jsme ještě nepomysleli. Kdybyste se na největší hrozby pro civilizaci zeptali v roce 1900, jaderné zbraně, genetické inženýrství

nebo umělá inteligence by nejspíš nikomu na mysl nepřišly, protože ještě nebyly vynalezeny. Co se týče následujícího století, jsme možná v téže situaci. Budoucí „neznámé neznámé“ mohou představovat větší hrozbu než rizika, o kterých víme dnes.

Kdykoli vynalezneme novou technologii, je to trochu jako sázet v ruletě proti jedinému číslu. Většinou vyhrájeme a technologie je celkově přínosná. Vždy ale existuje malá pravděpodobnost, že nám poskytne víc ničivé síly, než dovedeme zvládnout, a přijdeme o všechno.



Každá nově vyvinutá technologie přináší nejen nebyvalý potenciál, ale také nebezpečí.

Jaké je úhrnné riziko vyhynutí lidstva se vším všudy?

Podle odhadu mnohých odborníků a odbornic, kteří se těmito otázkami zabývají, je pravděpodobnost vymření lidstva v následujícím století mezi 1 a 20 %.

V podcastu hovoříme s Willem McAskillem o tom, proč je podle něj riziko v tomto století 1 %⁶.

Toby Ord v knize *Nad propastí: Existenční riziko a budoucnost lidstva* udává odhad, že celkové existenční riziko v tomto století je 1 : 6 (tedy 17 %) – jako při hodu kostkou. (Je ale třeba vzít v úvahu, že existenční katastrofa podle Ordovy definice není totéž, co vyhynutí lidstva – spadá sem i například celosvětová katastrofa, ze které by se náš druh už nikdy úplně nevzpamatoval, ačkoli by někteří lidé přežili.)

V knize uvádí následující tabulku s (velmi hrubými) odhady existenčních rizik počítaje těmi, která považuje za nejzávažnější.

6 <https://80000hours.org/podcast/episodes/will-macaskill-century-in-a-decade-navigating-intelligence-explosion/>

Druh existenční katastrofy	Pravděpodobnost během příštích sto let
Dopad asteroidu či komety	~ 1 : 1 000 000
Supervulkanická erupce	~ 1 : 10 000
Hvězdná exploze	~ 1 : 1 000 000 000
Úhrnné přírodní riziko	~ 1 : 10 000
Jaderná válka	~ 1 : 1 000
Změna klimatu	~ 1 : 1 000
Jiné ničení životního prostředí	~ 1 : 1 000
„Přirozené“ vzniklá pandemie	~ 1 : 10 000
Uměle vytvořená pandemie	~ 1 : 30
AI nesladěná s lidskými hodnotami	~ 1 : 10
Nepředvídaná antropogenní rizika	~ 1 : 30
Jiná antropogenní rizika	~ 1 : 50
Úhrnné antropogenní riziko	~ 1 : 6
Úhrnné existenční riziko	~ 1 : 6

Tyto hodnoty jsou asi milionkrát vyšší, než si lidé obvykle myslí.

Jak bychom to měli chápat? Badatelky a badatelé zabývající se touto oblastí se těmto tématům věnují zřejmě jen proto, že je považují za velmi důležité. Proto není překvapivé, že jejich odhady (v důsledku výběrového zkreslení jsou vysoké). Ačkoli s těmito čísly lze polemizovat, řadu názorů (včetně těch, které zastávají MacAskill a Ord) považujeme za realistické.

Proč může pomoc při ochraně budoucnosti být to nejdůležitější, co byste v životě mohli dělat

Nakolik bychom měli snižování těchto rizik upřednostňovat před jinými oblastmi, jako je např. světová chudoba, vymýcení rakoviny nebo změna politiky?

Cílem výzkumu naší organizace 80,000 Hours je pomoci lidem při hledání profesní dráhy s kladným dopadem na společnost. Přitom se snažíme identifikovat nejpalčivější problémy na světě, kterými se lze zabývat. Hodnotíme je pomocí našeho postupu pro srovnávání problémů z následujících hledisek:

- Rozsah – kolika lidí se problém týká
- Opomíjenost – kolik lidí se problému už věnuje
- Řešitelnost – jak snadné je dosáhnout pokroku

Když tento postup použijete, podle nás vám vyjde, že ochrana budoucnosti je největší

světovou prioritou. Takže pokud chcete, aby vaše profesní směřování mělo značný pozitivní dopad, toto je hlavní oblast, na kterou se zaměřit.

V příštích několika oddílech toto téma posoudíme z hlediska rozsahu, opomíjenosti a řešitelnosti. Budeme přitom vycházet z článku Nicka Bostroma *Existential Risk Prevention as a Global Priority* (Prevence existenčních rizik jakožto světová priorita), nepublikované práce Tobyho Orda a z vlastního výzkumu.

Začněme rozsahem problému. Uvedli jsme, že pravděpodobnost našeho vyhynutí v následujícím století je víc než 3 %. Jak velký je to problém?

Jedním z údajů, které můžeme zkoumat, je, kolik lidí by při takovém neštěstí zemřelo. V polovině století bude na Zemi asi 10 miliard obyvatel, takže při 3% pravděpodobnosti, že všichni zahynou, je střední počet obětí asi 300 milionů. To je pravděpodobně více než očekávaný počet úmrtí během následujících sta let na nemoci spojené s chudobou, jako je malárie.

Mnohá rizika, která jsme popsali, by také nemusela způsobit zánik civilizace, ale „střední“ katastrofu, což je asi pravděpodobnější. Z dříve zmíněného průzkumu vychází 10% pravděpodobnost katastrofy v následujícím století, která by si vyžádala více než 1 miliardu obětí. To by znamenalo nejméně dalších 100 milionů následných úmrtí a dalšího utrpení přeživších.

Takže i kdybychom vzali v úvahu jen dopad na současnou generaci, mezi nejzávažnější problémy, kterým lidstvo čelí, tato rizika katastrofy patří.

Rozsah problému je tímto ale značně podhodnocen, protože kdyby zanikla civilizace, přijdeme také o celou budoucnost.

Většina lidí chce zanechat vnoučatům lepší svět a je toho názoru, že bychom měli mít na paměti i budoucí generace obecně. V budoucnu může skvěle žít mnohem víc lidí, než je naživu dnes, a měli bychom brát ohled i na jejich zájmy. Je možné, že lidská civilizace přetrvá miliony let, takže když vezmeme v úvahu dopad rizik na budoucí generace, v sázce je milionkrát víc – v dobrém i ve zlém. Jak napsal Carl Sagan v časopise *Foreign Affairs* o ceně za jadernou válku⁷:

„Jaderná válka ohrožuje všechny naše potomky po celou existenci lidstva. I když počet obyvatel zůstane stejný, při průměrné délce života asi 100 let a dosažení doby obvyklé pro biologickou evoluci úspěšného druhu (zhruba 10 milionů let) přijde ještě asi 500 bilionů lidí. Podle tohoto kritéria je pro případ vyhynutí v sázce milionkrát víc než pro případ skromnější jaderné války, která by měla mít „jen“ stovky milionů obětí. Pro možnou ztrátu existuje mnoho dalších měřítek – včetně kultury a vědy, evoluční historie planety a významu životů všech našich předků, kteří přispěli k budoucnosti svých potomků. Vymření by znamenalo zkázu celého lidského konání.“

Jsme rádi, že Římané lidstvo vyhnout nenechali, protože to znamená, že může existovat celá moderní civilizace. Podle našeho názoru dlužíme tuto odpovědnost těm, kteří přijdou po nás – za předpokladu, že (jak věříme) budou žít naplňující život. Bylo

7 <http://www.jstor.org/stable/20041818>

by bezohledně a nespravedlivě ohrozit jejich existenci jen proto, aby se nám nakrátko dařilo lépe.

Nejde jen o to, že budoucnost se může týkat více lidí. Jak Sagan také zdůraznil, ať si vážíte čehokoli, nejspíš toho v budoucnu bude mnohem víc. Budoucí civilizace by mohla vytvořit svět bez nutnosti války a dosáhnout nevidaných duševních a uměleckých počinů. Mohli bychom vytvořit mnohem spravedlivější a ctnostnější společnost. A neexistuje žádný zásadní důvod, proč by lidé nemohli dosáhnout jiných planet, kterých je v naší galaxii asi 100 miliard. Pokud však způsobíme zánik civilizace, k ničemu z toho nikdy nebude moci dojít.

Nejsme si jistí, že tato skvělá budoucnost skutečně nastane – ale možnost to zjistit je další důvod, proč civilizaci zachovávat. Nepředat pochodeň další generaci je možná to nejhorší, čeho bychom se mohli kdy dopustit.

Asi největším problémem, kterému dnes svět čelí, se tudíž zdá být právě toto několikaprocentní riziko kolapsu civilizace. A je zarážející, jak moc tuto hrozbu zanedbáváme.

Proč tato rizika patří mezi nejopomíjenější světové problémy

Zde je přehled, kolik prostředků ročně plyne do některých důležitých oblastí:

Oblast	Roční cílené výdaje z veškerých zdrojů (velmi přibližně)
Věda a vývoj celosvětově	1,5 bilionu \$
Luxusní zboží	1,3 bilionu \$
Sociální zabezpečení v USA	900 miliard \$
Změna klimatu	> 300 miliard \$
Pro chudé na celém světě	> 250 miliard \$
Jaderná bezpečnost	1–10 miliard \$
Prevence extrémní pandemie	1 miliarda \$
Výzkum bezpečnosti AI	10 milionů \$

Jak vidíte, vynakládáme velké množství zdrojů na vědu a výzkum, abychom vyvinuli ještě mocnější technologii. Hodně také utratíme při (možná pomýlené) snaze zlepšit si život nakupováním luxusního zboží.

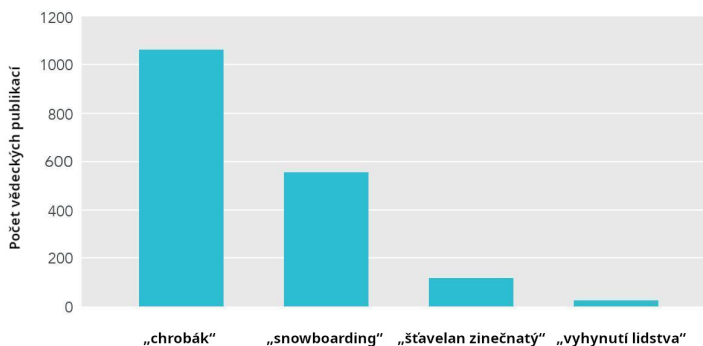
Mnohem méně jde na zmírňování rizika katastrofy způsobené změnou klimatu. Jen USA na sociální zabezpečení vydává několiknásobně více prostředků, než kolik do zmírňování dopadů změny klimatu investuje celý svět.

Přesto na ni ve srovnání s jinými hrozbami, které jsme popisovali, plynou ohromné prostředky. Náš hrubý odhad je, že na prevenci extrémní světové pandemie se vynakládá 300krát méně, ačkoli míra rizika se zdá přibližně podobná.

Nejopomíjenější je výzkum s cílem vyhnout se nehodám způsobeným AI. Plyne na něj nejspíš ještě 100násobně méně prostředků, přibližně jen 10 milionů \$ ročně.

Podobný obrázek se naskytne při porovnání nikoli vynaložených financí, ale počtu lidí, kteří se rizikům věnují. Získat údaje o penězích je ale snazší.

V případech vědecké pozornosti vidíme podobné zanedbávání (ačkoli některá jednotlivá rizika se těší významné pozornosti, například změna klimatu):



A domníváme se, že při uvážení politické pozornosti budou výsledky podobné jako u financování. Její drtivá většina připadá na konkrétní oblasti, které krátkodobě pomohou současné generaci, protože to přináší hlasy. Rizika katastrofy se berou v úvahu mnohem méně. Nejvíce pozornosti mezi nimi se pak dostává změně klimatu, zatímco nejvíce opomíjené jsou oblasti jako pandemie a AI.

Zanedbávání z hlediska zdrojů, výzkumu i politické pozornosti je při uvážení ekonomických zásad, na kterých jsou tyto věci založené, zcela očekávatelné. Z toho důvodu také tato oblast skýtá příležitosti lidem, kteří chtějí zlepšit svět.

Za uvedená rizika předně nemá odpovědnost žádný konkrétní stát. Představte si, že by Spojené státy významně zainvestovaly do prevence změny klimatu. Těžili by z toho všichni na světě, v USA ovšem žije jen 5 % světové populace, takže tamní občané by měli z přínosu těchto vynaložených prostředků jen 5 %. To znamená, že USA do této oblasti investují zoufale málo na to, jak prospěšná pro svět je. A totéž platí pro všechny ostatní země.

Řešením by byla spolupráce všech – kdyby každý stát ke zmírnění změny klimatu proporčně přispěl, všechny státy by se vyhnuly nejhorším důsledkům změny, a tím by z toho těžily.

Z pohledu každého jednotlivého státu je ovšem bohužel lepší, když své emise sníží

všechny ostatní státy, zatímco jeho vlastní hospodářství zůstane bez omezení. Každý stát má tudíž motivaci klimatické dohody porušovat, a proto je pokrok tak mizivý (jde o případ věžňova dilematu).

A i tohle je ve skutečnosti výrazné zlehčení problému. Největší prospěch ze snižování hrozeb katastrofy případně budoucím generacím. A ty nemají své zájmy jak hájit – ekonomicky ani politicky.

Kdyby se mohly účastnit našich voleb, z drtivé většiny by hlasovaly pro bezpečnější strategie. Podobně tak kdyby mohly posílat zpět v čase peníze, byly by ochotné věnovat nám za snížení rizik ohromné částky. (Snižování zásadních rizik vytváří technicky vzato mezigenerační celosvětové veřejné dobro, takže jde nejspíš o nejzanedbávanější způsob konání dobra.)

Náš současný systém v ochraně budoucích generací selhává. Známe lidi, kteří mluvili s nejvyššími vládními činiteli a činitelkami Velké Británie. Mnozí tito politici by rizika rádi řešili, avšak zaměřovat se na ně je kvůli tlaku mediálního a volebního cyklu podle nich náročné. Většina států nemá orgán, do jehož pravomoci by mírnění těchto rizik přirozeně spadalo.

Tato situace je sklíčující, ale zároveň představuje příležitost. Těm, kteří chtějí svět opravdu zlepšit, tento nedostatek pozornosti skýtá spoustu možností, jak pomoci, aby to mělo velký dopad.

Kapitola 6

Prevence katastrofálních pandemií

Cody Fenwick a tým 80,000 hours / 2020, aktualizováno 2025

Epidemie jsou nevyhnutelné; pandemie jsou volitelné. — Larry Brilliant

Mezi nejvíce smrtící události v dějinách patří pandemie. Covid-19 ukázal, že jsou hrozbou i dnes, a budoucí nákazy by mohly mít smrtelnost mnohem větší.

V důsledku technologického rozvoje nám ve skutečnosti hrozí horší biologické katastrofy než kdy dřív.

Pravděpodobnost katastrofální pandemie natolik závažné, že by ochromila civilizaci a ohrozila budoucnost lidstva, se zdá nepříjemně vysoká. Podle nás patří toto riziko mezi nejnaléhavější problémy světa.

Existuje zároveň řada praktických možností, jak *biologická existenční rizika* (GCBRs) snižovat. Jsme tudíž toho názoru, že jejich zmírňování je v současnosti jedním z nejslibnějších způsobů, jak chránit budoucnost lidstva.

Proč své profesní směřování věnovat prevenci vážných pandemií?

Při covidu-19 vyšlo jasně najevo, že jsme vůči světovým pandemiím zranitelní, a ukázaly se slabiny naší schopnosti reagovat. Přes pokroky v medicíně a hygieně bylo celosvětově zaznamenáno přibližně 7 milionů úmrtí a podle mnohých odhadů bylo obětí mnohem více¹.

Události jako morová rána ve 14. století a španělská chřipka svědčí o tom, že pandemie mohou pro lidstvo patřit mezi nejničivější katastrofy. Padají jim za oběť desítky milionů lidí, tedy znatelná procenta populace.

Představa možného dopadu patogenu nakažlivějšího a smrtelnějšího, než s jakými jsme se doposud setkali, je děsivá.

V principu by se takový patogen bohužel objevit mohl, zejména s ohledem na pokroky v biotechnologiích. Vědci a vědkyně dokáží navrhnout a vytvořit biologické činitele snáze a přesněji než dříve. Vytvoření patogenu, který by mohl ohrozit celé

Původně vyšlo jako *Preventing catastrophic pandemics* na

80000hours.org/problem-profiles/preventing-catastrophic-pandemics/. Zde zkráceno.

1 <https://www.nature.com/articles/d41586-022-04138-w>

lidstvo, bude s pokrokem v tomto oboru možná čím dál snazší.

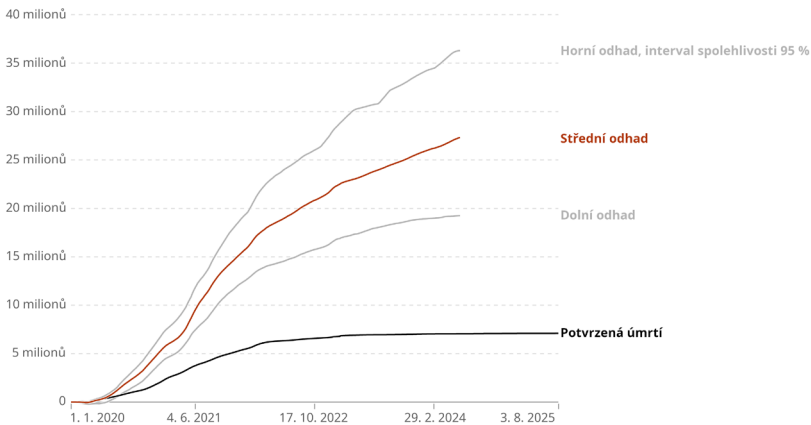
Státy nebo zločinci, kteří by k takovým patogenům měli přístup, by je mohli použít jako zbraň nebo tím vyhrožovat, aby získali výhody.

Při bezpečnostní chybě by také nebezpečné patogeny vytvořené k vědeckým účelům mohly omylem uniknout z laboratoře.

Odhad kumulativních nadúmrtí během covidu-19, celosvětově

Our World
in Data

U zemí, které nezveřejňovaly za daný týden údaje o úmrtnosti ze všech příčin, je uveden odhad s intervalem spolehlivosti. Pokud jsou zveřejněná data dostupná, je uvedena pouze hodnota. Pro srovnání jsou uvedena potvrzená kumulativní úmrtí na covid-19.



Data source: The Economist (2024); World Health Organization (2025)

OurWorldinData.org/coronavirus | CC BY

Jak vysvětlíme níže, obě situace by mohly vést ke katastrofální „uměle vytvořené pandemii“, která by podle nás mohla lidstvo ohrozit ještě více než pandemie vzniklé přirozeně.

Snaha využít nákazu jako zbraň naštěstí není příliš rozšířená, a ti, kdo by takový útok vykonat chtěli, by možná neusilovali o vytvoření neškodlivějšího možného patogenu. Úhrnná možnost nehody, bezohlednosti, zoufalství a výjimečné krutosti však znamená, že pravděpodobnost vypuštění patogenu, který by zabil vysoké procento populace, je znepokojivě vysoká. Riziko by bylo zvláště vysoké při konfliktu velmocí.

Může ale uměle vytvořená pandemie představovat hrozbu, že lidstvo vyhyne?

Názory se z dobrých důvodů různí. V minulosti se společnosti vzpamatovaly z pandemií, které vyhladily až 50 % obyvatelstva, možná i více.

Jsmo ale toho názoru, že v tomto století by pandemie mohly patřit mezi největší zdroje existenčního rizika. Zdá se, že biologie brzy pokročí natolik, že bude možné vyvolat pandemie, které by mohly vyhubit více než 50 % populace – a to ne jen na určitém místě, ale celosvětově. Není vyloučeno, že budou natolik závažné, aby způsobili

vyhynutí lidstva nebo přinejmenším škody, ze kterých se civilizace už nevzpamatuje.

Zdá se tedy, že je velmi důležité snižovat rizika biologických katastrof budováním pojistek proti jejich vypuknutí a přípravou na zmírňování jejich nejhorších dopadů.

Není zřejmě obvyklé, že by se lidé v širším odvětví biologické ochrany a připravenosti na pandemie věnovali přímo snižování katastrofických rizik a uměle vytvořených pandemií. Na projekty zmírňující rizika biologické katastrofy také připadá relativně malé procento prostředků určených na zdravotní bezpečnost.

Škody z biologických katastrof podle nás se vzrůstající závažností stoupají nelineárně, protože se zvyšuje možnost, že událost přispěje k existenčnímu riziku. Vyplyvá z toho, že především projektům zaměřeným na prevenci nejzávažnějších scénářů by se mělo dostávat více prostředků a pozornosti než v současnosti.

Ve zbytku tohoto oddílu se budeme zabývat srovnáním rizik přirozených a uměle vytvořených pandemií. Později se zaměříme na to, jakým činnostem se lze a je třeba pro snížení těchto rizik věnovat.

Vytvořili jsme také přehled profesí v oblasti výzkumu biologických rizik, strategií a politiky. Najdete tu konkrétnější informace o činnostech s velkým dopadem, kterými se lze zabývat, a rady, jak se do oboru dostat.

Přirozené pandemie svědčí o ničivosti biologických hrozeb

Čtyři nejhorší pandemie známých dějin byly tyto:

1. *Justinánský mor* (541–542 n. l.) pravděpodobně propukl v Asii a rozšířil se do Byzantské říše a Středomoří. První vlna měla nejspíš na svědomí kolem 6 milionů životů (asi ~ 3 % světové populace) a přispěla ke zvrácení byzantských územních zisků.
2. *Černá smrt* („první morová rána“) (1335–1355) zahubila podle odhadů 20–75 milionů lidí (asi 10 % obyvatelstva planety) a pravděpodobně měla dalekosáhlé důsledky pro průběh evropských dějin.
3. *Kolumbovská výměna* (16. stol.) byl sled pandemií, nejspíš včetně pravých neštovic a paratyfu, zavlečených evropskými kolonisty do Ameriky, kde pak devastovaly místní obyvatelstvo. Nejspíš hrály roli ve vymření přibližně 80 % původního obyvatelstva Mexika v 16. století. Mezi dalšími obyvateli v obou Amerikách měly na svědomí devastaci ještě většího procenta místních společností. V některých společnostech na tyto nemoci zahynulo až 98 % obyvatel.
4. *Španělská chřipka* (1918) se rozšířila téměř po celém světě a vyžádala si 50–100 milionů životů (2,5 až 5 % světové populace). Měla možná ještě více obětí než první nebo druhá světová válka.

Tyto historické pandemie ukazují, jak ničivé mohou biologické hrozby být a že samy o sobě představují riziko, které má smysl zmírňovat. Vidíme na nich také, že mimořádně ničivá pandemie může zapříčinit klíčové prvky světové katastrofy, mezi které

patří vysoká procentní úmrtnost a kolaps civilizace.

Přes tyto hrůzy minulosti se ale s ohledem na to, co o událostech přírodní historie víme, zdá nepravděpodobné, že by přírodní pandemie mohla být natolik závažná, aby sama v blízké budoucnosti způsobila vyhynutí celého lidstva.

Jak uvádí filozof Toby Ord v kapitole o přírodních rizicích v knize *Nad propastí*, z dějin vyplývá, že základní riziko vyhynutí pro lidstvo – pravděpodobnost, že nepřechká běžné situace – v důsledku přírodních příčin v příštích zhruba sto letech je velmi nízké.

Pokud by totiž základní riziko bylo přibližně 10 % za století, znamenalo by to, že po přibližně 200 tisíc let existence lidstva jsme měli dost štěstí. Mnohem méně překvapivé by to, že existujeme, bylo, kdyby toto riziko bylo kolem 0,001 % za století.

Žádná z nejhorších známých nákaz v dějinách nestačila k rozvrácení civilizace na celém světě natolik, že by to ohrozilo budoucnost našeho druhu. A obecně platí, že vyhynutí druhů následkem působení patogenů je v přírodě velmi vzácné.

Má riziko v důsledku přírodní pandemie klesající, nebo stoupající tendenci?

Ohrožují nás přírodní pandemie méně než dříve, nebo nás naopak pokrok lidské společnosti vystavuje riziku více?

Kvalitní data se hledají těžko. Zátěž na lidskou společnost v podobě infekčních nemocí má obecně klesající tendenci, to nám ale neříká mnoho o tom, jestli se nezhoršují zřídka výskyty masových pandemií.

Čistě teoreticky existuje mnoho důvodů pro pokles rizika. Například:

- Máme lepší hygienu než v minulosti a nejspíš se bude zlepšovat dále.
- Dokážeme vyrobit účinné očkovací látky a léky.
- Lépe rozumíme přenosu chorob, infekcím a jejich působení na organismus.
- Lidská populace je celkově zdravější.

Na druhou stranu:

- Mezinárodní obchod a letecká doprava umožňují, aby se nemoci šířily rychleji a dále. Létání například zřejmě významně napomáhalo mezinárodnímu šíření covidu-19. V minulosti byla ohniska nákaz kvůli náročnosti cestování do vzdálených míst nejspíš lokálnější.
- Kvůli změně klimatu se může zvýšit pravděpodobnost zoonotických onemocnění.
- Větší hustota obyvatel může zvýšit pravděpodobnost rychlého šíření chorob.
- Z mnohem větších populací domácích zvířat by se mohla přenášet onemocnění na lidi.

A nejspíš existují i další významné skutečnosti. Podle našeho odhadu frekvence přírodních pandemií vzrůstá, průměrně jsou ale méně závažné. Odhadujeme také, že druhý faktor je významnější než ten první, takže ve výsledku se celkové nebezpečí snižuje. Mnoho otázek ale zůstává nezodpovězených.

Uměle vytvořené patogeny by mohly být ještě nebezpečnější

Přestože však rizika představovaná přírodními pandemiemi klesají, rizika uměle vytvořených patogenů jsou téměř jistě na vzestupu.

Kvůli technologickému pokroku je totiž čím dál více možné vytvářet nebezpečné viry a infekční činitele. Nehoda nebo zneužití této technologie představuje věrohodné riziko celosvětové katastrofy, která by mohla ohrozit budoucnost lidstva.

Jedním z možných scénářů je, že by nějaký zločinec chtěl znovu vyvolat katastrofální historické pandemie.

Zcela uměle byly znovu vytvořeny viry dětské obrny, španělské chřipky a nedávno také koňských neštovic (což je blízký příbuzný pravých neštovic). Genetické sekvence všech těchto patogenů jsou veřejně dostupné a pokrok v biotechnologii a její rostoucí rozšířenost skýtá desíky možností.

Kromě obnovení historických chorob by mohla pokročilá biotechnologie vést také k tomu, že někdo vytvoří patogen nebezpečnější než ty, které v minulosti vznikly přirozeně.

Při vývoji virů neprobíhá jejich přirozený výběr tak, aby byly co nejničivější a nejvíce smrtící. Pokud by ale někdo chtěl ubližovat záměrně, mohl by zkombinovat ty nejhorší aspekty virů tak, jak by k tomu v přírodě velmi pravděpodobně nedošlo.

Sekvenování, úprava a syntéza DNA je dnes možná a čím dál snazší. Přibližujeme se tomu, že budeme schopni vytvářet biologické činitele podobně, jako navrhujeme a vyrábíme počítače nebo jiné produkty (ačkoli nelze říct, kdy k tomu dojde). To by lidem mohlo umožnit vytvářet patogeny smrtelnější, snáze přenosné nebo možná s úplně novými vlastnostmi.²

Vědci a vědkyně zjišťují, co způsobuje vyšší nebo nižší smrtelnost a nakažlivost patogenů, což umožňuje snáze předcházet propuknutí epidemií a zmírňovat je.

Zároveň to ale znamená, že informace k výrobě nebezpečnějších patogenů jsou čím dál dostupnější.

Všechny tyto technologie mají kromě rizik také možná medicínská využití. Virové inženýrství se například využívá v genové terapii a u vakcín (včetně některých proti covidu-19).

Znalost manipulace s viry, která umožňuje vyrábět lepší vakcíny nebo léky, lze ovšem zneužít ke „zlepšování“ biologických zbraní. Správné nakládání s těmito poznatky vyžaduje velkou obezřetnost.

Náznamy těchto nebezpečí lze nalézt ve vědecké literatuře. Experimenty na chřipce při výzkumu zisku funkce ukázaly, že umělým výběrem lze vytvořit patogeny s vlastnostmi, které zvyšují jejich nebezpečnost.

A vědecká obec ještě nezavedla dostatečně přísná pravidla, která by odrazovala od neomezeného sdílení nebezpečných objevů – jako například jak zvýšit smrtelnost

2 <https://tinyurl.com/27z29xcn>

virů – a předcházela mu. Proto zájemce o tento obor upozorňujeme, že biologická bezpečnost zahrnuje informační rizika. Je důležité, aby lidé, kteří s nimi pracují, byli prozíraví.

Vědci a vědkyně mohou v laboratoři učinit nebezpečné objevy nezáměrně. Při výzkumu vakcíny lze například odhalit mutace viru, které vedou k větší nakažlivosti choroby. A v jiných oblastech biologie, jako je například výzkum enzymů, vidíme, jak může naše pokročilá technologie dát vzniknout novým, možná nebezpečným schopnostem, které se v přírodě nikdy nevyskytly.

Ve světě s mnoha „neznámými neznámými“ možná objevíme mnoho nových nebezpečí.

Ačkoli tedy vývoj vědy přináší velké pokroky, otevírá také zločincům možnost záměrného vytváření nových nebo modifikovaných patogenů. Přestože se převážná většina vědecké obce snaží lidstvu prospět, mnohem menší skupina může prostřednictvím těchto poznatků napáchat velké škody.

Pokud by někdo měl dostatečnou motivaci, zdroje a technické dovednosti, je těžké si domyslet, jak katastrofální umělou pandemii by jednou mohl způsobit. S pokrokem technologií budou nástroje k vyvolání biologické katastrofy čím dál dostupnější a práh pro dosažení hrozivých výsledků se může stále snižovat – což bude zvyšovat riziko velkého útoku. K riziku může přispět zejména pokrok AI.

Hrozbu představuje zneužití i nehoda

Rizika uměle vytvořené pandemie můžeme rozdělit na omyl a zneužití – stručně řečeno si lze na jedné straně představit nepodařený vědecký experiment a na druhé bioteroristický útok.

Historie nehod a úniků z laboratoří, kvůli kterým byli lidé vystaveni nebezpečným patogenům, nahání husí kůži:

- V roce 1977 se objevil neobvyklý kmen chřipky, která se neúměrně projevovala u mladých lidí. Zjistilo se, že geneticky pochází z kmene, jehož vývoj se zastavil v roce 1950. To naznačovalo laboratorní původ z nepovedeného testu vakcíny.
- V roce 1978 vedl únik viru z laboratoře ve Velké Británii k poslednímu úmrtí na pravé neštovice.
- V roce 1979 unikly z laboratoře v Sovětském svazu, kde se nejspíš vyráběly biologické zbraně, spóry anthraxu. Choroba se šířila městem, napadala obyvatel i zvířata a vyžádala si asi 60 obětí. Přestože se událost zprvu tutlala, ruský prezident Boris Jelcin později přiznal, že infekce omylem unikla vzduchem z vojenské laboratoře.
- V roce 2014 byly desítky pracovníků Střediska pro kontrolu a prevenci nemocí v USA potenciálně vystaveny živému anthraxu, když vzorky, které měly být neaktivní, nebyly usmrceny dostatečně a byly rozeslány do laboratoří nižších

úrovni, kde se ne vždy používalo vhodné ochranné vybavení.

- Nevíme přesně, jak často k takovým událostem dochází, protože nejsou důsledně zaznamenávány. A v mnoha dalších případech možná chybělo málo.

V minulosti také došlo k mnoha teroristickým útokům a vývoji zbraní hromadného ničení ze strany států. Mezi bioteroristické a biologické útoky patří například tyto:

- V roce 1763 dala britská armáda v pevnosti Fort Pitt pokrývky z ošetrovny, kde se léčily neštovice, místním indiánským kmenům, aby tak chorobu rozšířila a obyvatele oslabil. Není jisté, zda to mělo kýžený účinek, ale mnohé kmeny byly neštovicemi zpusťošeny.
- Japonská jednotka 731 během 2. světové války páchala v Číně děsivé pokusy na lidech a vedla biologickou válku. Pozabíjela tisíce lidí a možná i mnohem více anthraxem, cholerou a morem. Podrobnosti vyšly najevo až později.
- V 60. a 70. letech probíhal v Jižní Africe tajný vládní program na vývoj chemických a biologických zbraní známý jako Project Coast. Jeho cílem bylo vyvinout biologické a chemické činitele namířené proti konkrétním etnickým skupinám a politickým oponentům. Součástí byl také vývoj látek působících neplodnost.
- Stoupenci hnutí Bhagvána Radžníše infikovali v roce 1984 salátové bary v oregonských restauracích salmonelou, kterou se tak nakazilo více než 750 osob. Důvodem byla snaha ovlivnit nadcházející volby.
- V roce 2001 krátce po útocích z 11. září byly v USA poštou rozeslány spóry antraxu do několika zpravodajských médií a dvěma senátorům. Nakazilo se 22 osob a pět zemřelo.

Měli bychom se tedy víc bát nehod, nebo bioterorismu? Těžko říct. Není k tomu k dispozici dostatek údajů a argumenty existují pro obojí.

Může se zdát, že znepokojivější je záměrné šíření smrtícího patogenu. Jak jsme uvedli, nejhorší pandemie by pravděpodobně byly spíše záměrně vytvořené než důsledek omylu. Navíc existují způsoby, jak škodlivost infekce snížit nebo zvýšit, a nezáměrný únik by pravděpodobně nebyl optimalizovaný tak, aby uškodil co nejvíc.

Naopak lze říct, že mnohem více lidí to myslí dobře a nechtějí prostřednictvím biotechnologie světa ubližovat, nýbrž prospět. Aktivita s cílem omezit státní programy na vývoj biologických zbraní zároveň nejspíš snižují počet možných útočníků. Zdá se tudíž pravděpodobné, že k tragické nehodě se naskytá víc příležitostí než k masovému zločinnému biologickému útoku.

Tyto argumenty jsou podle našeho odhadu v souhrnu těmi nejvýznamnějšími faktory. Máme tedy za to, že záměrné zneužití je nebezpečnější než náhodný únik, ačkoli je rozhodně třeba bránit obojímu.

Obecné rizikové faktory: srovnání případů

	Přírodní patogen	Uměle vytvořený kmen
Uniklý omylem nebo přirozeně se vyskytující	Nízké riziko	Někde mezi
Záměrné rozšíření	Někde mezi	Vysoké riziko

*Nemám silný názor na relativní riziko v pravém horním a levém spodním kvadrantu, ale očekávané riziko není stejné nebo podobné

Celkově se toto riziko jeví jako významné

Setkali jsme se s různými odhady rizika biologické existenční katastrofy včetně možnosti uměle vytvořené pandemie. Asi nejlepší odhady padly ve forecasterském turnaji Existential Risk Persuasion Tournament (XPT).

Projekt spočíval v tom, že skupiny odborníků a odbornic z oboru a zkušených forecasteřů a forecasterek měly odhadovat pravděpodobnost extrémních událostí. Mediánové odhady pravděpodobnosti biologických rizik uváděné forecastery a odborníky vypadaly následovně:

- Katastrofa (událost, která způsobí vyhynutí 10 % lidské populace nebo více) do roku 2100: ~1–3 %
- Událost, která způsobí vyhynutí lidstva: 1 : 50 000 až 1 : 100
- Geneticky vytvořený patogen vyhubí do roku 2100 více než 1 % populace: 4–10 %
- Poznámka: forecasteři obvykle udávali nižší odhady rizik než odborníci z oboru.

Přestože jde o nejlepší údaje, se kterými jsme se setkali, lze k nim mít spoustu výhrad.

Hlavní tři jsou tyto:

1. Existuje málo důkazů o tom, že by něčí dlouhodobé prognózy mohly být přesné. Úspěšnost minulých prognóz se hodnotí podle otázek, které se vyřeší během měsíců nebo let, nikoli desetiletí.
2. Odhady se mezi jednotlivými skupinami i v rámci nich velmi lišily – účastníci udávali čísla rozdílná mnohonásobně nebo i o řády.
3. Byli vybráni odborníci, kteří se rizikům katastrofy již věnují – odhad extrémních rizik udaný typickým odborníkem v některých odvětvích hygieny by mohl být obecně nižší.

Těžko s jistotou říct, jak tyto různé odhady a argumenty posuzovat, a považujeme za pochopitelné, že lidé docházejí různým závěrům.

S ohledem na to, jak závažná by katastrofální pandemie byla, že možná ničivá síla

uměle vyvolané pandemie se zdá skoro neomezená a že opatření na její zmírnování jsou široce prospěšná, jsme ale toho názoru, že by se tomuto problému mělo věnovat mnohem víc lidí než dnes.

Snižování rizik biologické katastrofy je hodnotné z hlediska mnoha různých pohledů na svět

Vzhledem k tomu, že naší prioritou jsou světové problémy, které by mohly mít významný dopad na budoucí generace, nejvíc nám záleží na práci na snižování nejzávažnějších biologických hrozeb – zejména těch, které by mohly způsobit vyhynutí lidstva nebo rozvrátit civilizaci.

Biologická bezpečnost a snižování rizika katastrofy ale představují oblasti s velkým dopadem pro lidi s různými světovázory, a to z těchto důvodů:

1. *Biologická katastrofa by ohrozila i krátkodobé zájmy.* Jak ukázal covid-19, rozsáhlé pandemie mohou vést k mimořádným nákladům pro lidi dnes. Ještě virulentnější nebo smrtelnější choroby by způsobily ještě větší počet úmrtí a víc utrpení.
2. *Opatření zmírňující největší biologická rizika umožní zároveň předcházet i běžnějším onemocněním.* Díky monitoringu onemocnění lze odhalit velká i malá ohniska, aktivity proti šíření nález umožňují zastavit závažnější i méně závažné záměrné zneužití, lepší osobní ochranné vybavení umožňuje předejít nejrůznějším infekcím atd.

Biologická bezpečnost se také značně překrývá s dalšími světovými problematikami, jako je třeba světové zdraví (např. Agenda celosvětové zdravotní bezpečnosti (GHSA)), velkochovy (např. iniciativy One Health) nebo AI.

Pro snížení těchto rizik můžeme podniknout jednoznačné kroky

Biologická bezpečnost a připravenost na pandemie jsou multidisciplinární oblasti. K efektivnímu řešení těchto hrozeb je zapotřebí celá řada aktivit včetně těchto:

- Badatelé v oblasti techniky a biologie by měli zkoumat a vyvíjet prostředky na potlačování nález
- Podnikatelé a praktičtí odborníci v odvětví by tyto prostředky měli rozvíjet a zavádět
- Výzkumníci v oblasti strategií a forecasteri by měli vytvořit plány
- Politici by měli schvalovat a zavádět opatření pro snížení biologických hrozeb

Konkrétně můžete:

- V oblasti řízení státu, akademické sféry, průmyslu nebo mezinárodních organizací usilovat o tyto cíle: Lepší řízení výzkumu zisku funkce, který zahrnuje práci s možnými pandemickými patogeny, lepší řízení komerční syntézy DNA a dalšího výzkumu a odvětví, která mohou napomoci k vytvoření zvlášť nebezpečných umělých patogenů (nebo rozšířit přístup k nim).
- Posilovat mezinárodní závazky, že státy nebudou vyvíjet nebo používat biologické

ké zbraně. Příkladem je Úmluva o zákazu biologických zbraní.

- Pracovat na vývoji nových technologií, které mohou zmírňovat nebo identifikovat pandemie a efekty biologických zbraní. Např.:
 - » Širokospektrální testování, léky a vakcíny – a jak je vyvíjet, vyrábět a šířit v krizové situaci
 - » Metody detekce – např. monitorování odpadních vod – které umožní identifikovat nová nebezpečná ohniska
 - » Jiné než farmaceutické prostředky, např. lepší osobní ochranné vybavení
 - » Další způsoby, jak zabránit přenosu vysoce rizikových onemocnění, jako např. dezinfekce dalekým ultrafialovým zářením far-UVC³
- Pracovat na zavádění a propagaci výše uvedených technologií, aby byla společnost chráněná a motivace k vyvolání pandemie se snižovaly
- Pracovat na zlepšení informační bezpečnosti, aby byl chráněný biologický výzkum, který by ve špatných rukou mohl být nebezpečný
- Věnovat se zkoumání, zda pokroky v AI biologická rizika nezhorší, a hledat možná řešení tohoto problému
- Další informace o prioritách biologické bezpečnosti najdete v našem článku o tom, jak podle odborníků nejlépe zvládnout příští pandemii.

Širší oblast biologické bezpečnosti a připravenosti na pandemie ke snížení rizik katastrofy významně přispívá. Mnohé dobré způsoby přípravy na pravděpodobnější, méně závažné nákazy sníží také ta nejhorší rizika.

Když například vyvineme širokospektrální vakcíny a léky na prevenci a léčbu široké škály možných pandemických patogenů, bude to všeobecně přínosné pro veřejné zdraví i biologickou bezpečnost. Zároveň to ale také nejspíš sníží pravděpodobnost nejhorších možných scénářů, které jsme uváděli – bioteroristický útok, který má způsobit katastrofu, není ve světě připraveném chránit se před nejpravděpodobnějšími nákazami tak snadný. A pokud stát nebo jiný aktér uvažující o výrobě takové zbraně ví, že svět je proti ní velmi pravděpodobně chráněný, má ještě menší důvod to vůbec zkoušet.

Podobně to platí pro lepší ochranné vybavení, určité formy monitorování nemocí a dezinfekci vzduchu ve vnitřních prostorách.

Pokud se ale chcete věnovat prevenci těch nejhorších scénářů, v oblasti biologické bezpečnosti a prevence pandemií existují aktivity, na které je lepší se zaměřit spíše než na jiné.

Podle některých odborníků v oboru, například biologa Kevina Esvelta z MIT, přispívá k nejlepším opatřením ke snižování rizika člověkem způsobených pandemií spíše fyzika a technika než biologie⁴.

Důvodem je, že každé biologické protiopatření, jako třeba očkování, nejspíš půjde

3 <https://tinyurl.com/2xw4th9c>

4 <https://tinyurl.com/2denx8zr>

biologickými prostředky obejít – jako se třeba viry mohou vyvíjet, aby imunitu získanou očkováním překonaly.

Schopnost nákazy překonat fyzická protiopatření ale může být omezená. Nejspíš například nelze vytvořit virus, který by dokázal proniknout dostatečně zabezpečeným osobním ochranným vybavením nebo přežít far-UVC záření. Pokud to skutečně platí, nejsilnější ochranou proti největším pandemickým hrozbám by mohla být opatření tohoto typu.

Kapitola 7

Umělá inteligence mění náš svět – je na nás všech, aby to dopadlo dobře

Max Roser / 2022

O budování AI dnes rozhoduje skupinka technologů. Vzhledem k tomu, že umělá inteligence mění naše životy, by mělo být v zájmu nás všech se o ní informovat a zabývat se jí.

Proč byste se měli zajímat o vývoj umělé inteligence?

Uvažte, jaké jsou jiné možnosti. Pokud si o AI vy a širší veřejnost nezjistíte podrobnosti a nebudete se v tomto tématu angažovat, necháme rozhodnutí, jak tato technologie změní svět, na hrstce podnikatelů a inženýrů.

To se děje v současnosti. Těchto pár lidí v několika technologických firmách, které se umělé inteligencí (AI) přímo věnují, si nesmírný rozsah její nadcházející moci uvědomuje¹. Pokud se nezapojí zbytek společnosti, právě tato hrstka elit rozhodne, jak AI naše životy promění.

Aby se současný stav změnil, rád bych v tomto článku odpověděl na tři otázky: Proč je obtížné brát vyhlídky na to, že AI změní svět, vážně? Jak si takový svět představit? A co je se vzrůstající mocí této technologie v sázce?

Proč vyhlídky na to, že AI změní svět, není snadné brát vážně?

Že technologie může podstatně proměnit svět, by mělo být jasné. Stačí se podívat, nakolik už k tomu došlo. Kdybyste s sebou, až příště někam poletíte, mohli do letadla

Původně vyšlo jako *Artificial intelligence is transforming our world — it is on all of us to make sure that it goes well* na ourworldindata.org/ai-impact

1 <https://ourworldindata.org/brief-history-of-ai>

vzít rodinu lovců a sběračů z doby před 20 tisíci lety, celkem by se divili. Technologie už náš svět proměnila, a měli bychom tedy očekávat, že se to může stát zase.

Přestože jsme se ale už s takovými proměnami setkali, probíhaly po několik generací. Dnešní situace je jiná v tom, že se nesmírně zrychlily. Technologie využívané v dětství našich předků pro ně bývaly zásadní i ve stáří. To už pro současné generace neplatí. Obvykle se spíš stává, že v pozdějším životě člověk běžně využívá technologie, které by byly v jeho mládí nepředstavitelné.

²To je první důvod, proč možnost proměny světa někdy bereme na lehkou váhu – rychlost, s jakou ji technologie může přinést, lze snadno podcenit.

Druhým důvodem, proč se představě transformativní AI – která by dosahovala až lidské inteligence – těžko přikládá význam, je, že poprvé jsme se s ní setkali v kině. Není divu, že na mnohé z nás působí myšlenka strojů s lidskými schopnostmi podobně jako budoucnost, kdy se po Zemi prohání upíři, vlkodlaci nebo zombie.³

Je ale dost dobře možné, že nejde jen o výplody sci-fi a fantasy, ale zároveň o zásadní vynález, který by mohl vzniknout za našeho života nebo života našich dětí.

Třetím důvodem, proč není snadné tuto vyhlídku brát vážně, je, že si nedokážeme představit, že by výkonná AI mohla způsobit dalekosáhlé změny. I to je pochopitelné. Utvořit si představu o budoucnosti velmi odlišné od naší současnosti není snadné. Existují dva pojmy, které pro tuto představu považují za nápomocné. Pojďme se na oba podívat.

Jak si budoucnost umělé inteligence představit?

Při úvahách o budoucnosti umělé inteligence považují za užitečné zohlednit především dva odlišné pojmy: AI lidské úrovně a transformativní AI.⁴ První pojem zdůrazňuje schopnosti AI ve srovnání se známým měřítkem, zatímco u transformativní AI jde zejména o vliv, jaký by tato technologie měla na svět.

Z našeho dnešního pohledu to může do značné míry znít jako sci-fi. Je proto dobré vzít v úvahu, že podle většiny dotázaných odborníků a odbornic na AI existuje reálná pravděpodobnost, že AI lidské úrovně vznikne během několika desetiletí, a podle některých i mnohem dříve.⁵

2 <https://ourworldindata.org/technology-long-run>

3 Ještě větší problém je představit si, jak by se budoucnost s AI lidské úrovně odvíjela.

Jakýkoli konkrétní scénář nebude zahrnovat jen myšlenku, že takto mocná AI existuje, ale i celou řadu dalších předpokladů o budoucích okolnostech, za jakých k tomu dojde. Je tudíž obtížné představit si scénář s AI lidské úrovně, který zároveň nepůsobí vykonstruovaně, pitvorně, nebo i směšně.

4 Oba tyto pojmy jsou ve vědecké literatuře o AI široce používané. Často se jejich pomocí například definují otázky časového rámce dalšího vývoje AI.

5 <https://ourworldindata.org/ai-timelines>

Výhody a nedostatky srovnávání strojové a lidské inteligence

O AI lidské úrovně se dá uvažovat mimo jiné tak, že ji porovnáme se stavem AI dnes. Zatímco dnešní systémy AI mají mnohdy podobné schopnosti jako určité omezené části lidské mysli, AI lidské úrovně by představovala stroj schopný vykonávat stejné spektrum mentálních činností jako my lidé.⁶ Byl by „schopen se naučit cokoli, co zvládne člověk“, jak uvádějí v učebnici o AI Norvig a Russell.⁷

Škála schopností charakterizujících inteligenci umožňuje lidem řešit problémy a dosahovat nejrůznějších cílů. AI lidské úrovně by tudíž představovala systém schopný řešit všechny problémy a vykonávat všechny činnosti, jaké dnes zvládáme my lidé. Takový stroj nebo soubor strojů by zastal práci překladatele, účetní, ilustrátorky, učitele, psychoterapeutky či řidiče kamionu a dovedl by obchodovat na světových finančních trzích. Stejně jako my by se také dokázal zabývat výzkumem a vědou a na základě toho vyvíjet nové technologie.

Koncept AI lidské úrovně má nesporné přednosti. Protože jako měřítko využívá naši inteligenci, kterou známe, nabízí nám jasné vodítko, jak si možnosti příslušné AI představit.

Tento pojem má ovšem i zjevné nedostatky. Když představu budoucích systémů AI vztahujeme ke známému konceptu lidské inteligence, vystavujeme se riziku, že se tím zastřou obrovské rozdíly mezi těmito skutečnostmi.

Některé z těchto rozdílů jsou nasnadě. Systémy AI budou například mít nesmírnou kapacitu počítačové paměti, vedle které naše schopnost uchovávat informace bledne. Další zjevný rozdíl spočívá v rychlosti stroje při přijímání a zpracovávání informací. Ukládání a rychlost zpracování dat ovšem nejsou jedinými rozdíly. Seznam oblastí, ve kterých stroje lidi předčí, neustále narůstá: v šachách například systémy AI dosáhly úrovně nejlepších lidských hráčů koncem 90. let a před více než deseti lety lidi překonaly. V méně vzdálené minulosti k tomu došlo například u deskové hry go a složitých

6 Lidé jsou schopni vykonávat spektrum duševních činností, což znamená, že podle toho, na jaký aspekt tohoto spektra se zaměříte, můžete dojít k různým definicím inteligence. (V článku o inteligenci na Wikipedii je například uvedena řada definic od různých vědců a v různých oborech.) Existují proto i různé definice „AI lidské úrovně“.

Vyskytuje se také několik úzce souvisejících pojmů. Termíny obecná umělá inteligence (Artificial General Intelligence), strojová inteligence vysoké úrovně (High-Level Machine Intelligence), silná AI (Strong AI) a plná AI (Full AI) se někdy používají synonymně a jindy mívají podobné, avšak samostatné definice. Pro konkrétní diskuse je ale nutné tento koncept určit úžeji. Přesnější definice AI lidské úrovně udávají pro účely svých konkrétních výzkumů například autoři a autorky prací o časových výhledech v oblasti AI.

7 Norvig, Peter a Russel, Stuart: Artificial Intelligence: A Modern Approach. 4. vyd. Pearson 2021.

strategií.^{8,9}

Tyto rozdíly by znamenaly, že AI, která by v každé oblasti byla přinejmenším tak dobrá jako člověk, by svou celkovou výkonností lidskou mysl dalece předčila. I první AI „lidské úrovně“ by tedy v mnoha ohledech lidi překonávala.¹⁰

Lidská inteligence je pro tu strojovou špatnou metaforou i v dalších ohledech. Naše uvažování se od toho strojového často velmi liší, a výsledek přemýšlení strojů nám tudíž může být velmi vzdálený.

Nejvíce udivující a znepokojivé jsou zvláštní a nečekané možnosti selhání strojové inteligence. Příkladem je generovaný obrázek koně, který vidíte níže. AI na jednu stranu dokáže něco, co žádný člověk nesvede: vytvořit obrázek čehokoli, v jakémkoli stylu za pouhých několik vteřin – a zároveň dělá chyby, kterých by se žádný člověk nedopustil.¹¹ Nikdo by koně omylem nenakreslil s pěti nohama.¹²

Představovat si mocnou AI budoucnosti prostě jako člověka by tudíž pravděpodobně byla chyba. Možná budou tyto systémy natolik odlišné, že připisovat jim „lidskou úroveň“ bude neodpovídající.

8 Systém AI AlphaGo a jeho různí nástupci zvítězili v go proti mistrům. Systém Pluribus lidi porazil v pokeru no-limit Texas hold ,em. Systém Cicero dokáže vítězit ve strategické hře Diplomacie, přičemž využívá strategii a lidský jazyk. Viz Meta Fundamental AI Research Diplomacy Team (FAIR), Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, et al.: Human-Level Play in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. *Science* 0(0), 22. 11. 2022. Dostupné z: <https://doi.org/10.1126/science.ade9097>

9 <https://ourworldindata.org/grapher/computer-chess-ability>

10 Zároveň je to problematické při posuzování, jak si inteligence stroje stojí ve srovnání s tou lidskou. Kdyby inteligence byla obecnou jednotlivou schopností, mohli bychom ji porovnávat a hodnotit snadno. Vzhledem k tomu, že ale jde o paletu dovedností, srovnává se strojová inteligence s lidskou mnohem hůře. Testy systémů AI proto sestávají z celé řady úloh. Viz např. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, Jacob Steinhardt (2020): Measuring Massive Multitask Language Understanding nebo definici kritérií pro obecnou umělou inteligenci v tomto odhadu na platformě Metaculus: <https://www.metaculus.com/questions/5121/date-of-artificial-general-intelligence/>

11 Přehled možných chyb systémů AI naleznete v článku Charles Choi: 7 Revealing Ways AIs Fail. Za pročetí stojí také AIAAIC Repository, kde jsou „podrobně uvedeny nedávné incidenty a polemiky vyvolané AI, algoritmy a automatizací nebo s nimi související.“

12 Tento příklad jsem si vypůjčil od výzkumníka AI Françoise Cholleta.



Obrázek koně vygenerovaný AI¹³

Transformativní AI je definována svými důsledky pro svět

Pojem transformativní AI naopak na srovnání s lidskou inteligencí založený není. Výhodou je, že se tak lze vyhnout problémům, které při srovnání vznikají. Na druhou stranu je obtížnější si představit, jak by takový systém vypadal a co by dokázal. Musíme se víc snažit. Je třeba představit si svět s inteligentními aktéry, kteří od nás mohou být velmi odlišní.

Transformativní AI není definována na základě konkrétních schopností, ale reálného dopadu, který by měla. Za transformativní lze podle badatelů považovat takovou AI, která „má dostatečnou moc, aby nám přinesla novou, kvalitativně odlišnou budoucnost.“¹⁴

V lidských dějinách došlo k takto významné proměně dvakrát: při zemědělské

13 Z Cholletových komentářů vyplývá, že byl vytvořen pomocí systému AI Stable Diffusion.

14 Jde o citaci Holdena Karnofského (2021) z článku AI Timelines: Where the Arguments, and the „Experts,“ Stand. Pro dřívější Karnofského úvahy o těchto pojmech v oblasti AI viz jeho článek Some Background on Our Views Regarding Advanced Artificial Intelligence z roku 2016.

Ajeya Cotra, o jejímž výzkumu časového rámce vývoje AI se zmiňují v jiných článcích této série, se snaží AI, kterou by bylo možné považovat za transformativní, definovat kvantitativně. Ve své často citované zprávě o časových rámcích vývoje AI (<https://tinyurl.com/284qy4sx>) ji definuje jako změnu v softwarové technologii, která zvedne míru růstu hrubého světového produktu „na 20 až 30 % ročně“. Několik dalších badatelů transformativní AI (TAI) definuje podobně.

a průmyslové revoluci.

Vznik transformativní AI by byl událostí podobného významu. Podobně jako nástup zemědělství před 10 tisíci lety nebo přechod od ruční výroby ke strojové by změnil život miliard lidí všude na planetě a celé směřování budoucnosti lidstva¹⁵.

Technologiím, které podstatně mění způsob produkce celé řady zboží a služeb, říkáme technologie obecného určení. Dvě předchozí transformativní události vyvolal objev dvou zvláště významných technologií tohoto typu. Šlo o změnu produkce potravin, kdy lidstvo přešlo z lovu a sběru na zemědělství, a vzestup strojové výroby při průmyslové revoluci. Na základě důkazů a argumentů v této sérii článků (¹⁶) o vývoji AI považuji za pravděpodobné, že podobně významnou technologií obecného určení by mohla být i mocná AI.

Časová osa tří transformativních událostí světových dějin:



Budoucnost AI lidské úrovni, nebo AI transformativní?

Tyto dva pojmy jsou úzce spjaté, ale nikoli totožné. Vytvoření AI lidské úrovni by na náš svět jistě mělo transformativní dopady. Pokud by práci většiny z nás mohla vykonávat AI, změnily by se životy milionů lidí.¹⁷

Opak však neplatí. Transformativní AI může vzniknout, aniž by se vyvinula AI lidské úrovni. Protože přirovnávat inteligenci strojů k lidské mysli je v mnoha ohledech nevhodné, je reálné, že transformativní AI vyvineme dříve než AI lidské úrovni. Může to s ohledem na daný vývoj také znamenat, že strojová inteligence, kterou by bylo užitečné přirovnat k lidské, se neobjeví nikdy.

Kdy a zda systémy AI některé z těchto úrovní dosáhnou, se pochopitelně předpovídá obtížně. V doprovodném článku na toto téma podávám přehled současných názorů badatelů a badatelek z oboru¹⁸. Podle mnoha z nich existuje reálná pravděpodobnost, že tyto systémy vyvineme v nadcházejících desetiletích, a podle některých k tomu dojde mnohem dříve.

15 <https://ourworldindata.org/the-future-is-vast>

16 <https://ourworldindata.org/artificial-intelligence>

17 AI lidské úrovni se obvykle definuje jako softwarový systém schopný vykonávat nejméně 90 nebo 99 % všech hospodářsky významných činností vykonávaných lidmi. Méně ambiciózně definováno by šlo o systém AI schopný vykonávat veškeré činnosti, které dnes může vykonávat člověk pracující distančně na počítači.

18 <https://ourworldindata.org/ai-timelines>

Co je s rostoucí mocí AI v sázce?

Všechny významné technologické inovace mají řadu kladných i záporných důsledků. V případě AI je spektrum možných dopadů – od těch nejnegativnějších k těm nejpozitivnějším – mimořádně široké.

Skutečnost, že tato technologie může působit škody, je zřejmá, protože k tomu již dochází.

Ublížovat tyto systémy mohou, když je lidé využívají se zlým úmyslem. Příkladem je využití AI k masovému dohledu nebo v politicky motivovaných dezinformačních kampaních.¹⁹

Škodit ale AI může i nechtěně – když se chová jinak, než bylo zamýšleno, nebo udělá chybu. V Nizozemsku například systém AI využívaný úřady nepravdivě uvedl, že odhadem 26 tisíc rodičů podává neoprávněné žádosti o příspěvky na dítě. Tato falešná obvinění způsobila mnoha nízkopříjmovým rodinám těžkosti, a nakonec vedla v roce 2021 k odstoupení vlády.²⁰

S rostoucí mocí AI se může rozsah negativních dopadů AI výrazně rozšiřovat. Mnohým těmto rizikům se dostává oprávněné pozornosti ze strany veřejnosti: výkonnější AI může vést k masovému nahrazování pracovních sil nebo ke krajní koncentraci moci a bohatství. Protože je AI vhodná k masovému dohledu a ovládnání obyvatelstva, v rukou autokratů by pak mohla posilovat totalitarismus.

Dalším extrémním rizikem je takzvaný *problém sladění hodnot AI s lidskými*. Jde o obavu, že mocný systém AI by nedokázal ovládat nikdo, i kdyby tento systém ubližoval lidem nebo škodil lidstvu jako celku. Tomuto riziku se bohužel širší veřejnost příliš nevěnuje, ačkoli podle předních badatelů a badatelek v oboru jde o mimořádně významnou hrozbu.²¹

19 Využitím AI v politicky motivovaných dezinformačních kampaních se zabývá např. John Villasenor (listopad 2020): How to deal with AI-enabled disinformation. Obecně k tomuto tématu viz Brundage a Avin et al. (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Dostupné z maliciousaireport.com. Vhodným rozcestníkem literatury a zpráv o masovém dohledu je příslušný článek na Wikipedii: https://en.wikipedia.org/wiki/List_of_government_mass_surveillance_projects

20 Viz například článek z Wikipedie o nizozemském skandálu souvisejícím s příspěvky na dítě (Dutch childcare benefits scandal) a článek Melissy Heikkilä (2022): Dutch scandal serves as a warning for Europe over risks of using algorithms na zpravodajském webu Politico. Technologie mohou také posilovat rasovou a genderovou diskriminaci. Viz kniha Briana Christiana *The Alignment Problem* a zprávy AI Now Institute (<https://ainowinstitute.org/publications>).

21 Přehled poskytují ve svých knihách Stuart Russell (2019): *Human Compatible* (hl. kap. 5; česky *Jako člověk*. Přel. Jiří Zlatuška. Praha: Argo, Dokořán 2021) a Brian Christian (2020): *The Alignment Problem*. Christian popisuje úvahy mnoha významných badatelů v oblasti AI od počátků až dosud a poskytuje vynikající přehled tohoto problému. Velký význam přičítají této hrozbě i některé přední soukromé společnosti, které na mocné AI pracují – viz článek

Jak by se AI mohla vymknout kontrole a ubližovat lidem?

Riziko nespočívá v tom, že by získala vědomí sebe sama, utvořila si zlé úmysly a „rozhodla se“ tak. Hrozí, že se AI pokusíme zadat určitý konkrétní cíl – i velmi chvályhodný – a ona začne lidem škodit při jeho sledování. Jde o problém nezamýšlených důsledků: AI bude dělat, co jsme jí řekli, nikoli to, co jsme od ní chtěli.

Nemůžeme jí tedy prostě říct, aby takové věci nedělala? Určitě jí lze vybudovat tak, že se předejde kterémukoli konkrétnímu problému, ale není snadné předvídat veškeré nezamýšlené škodlivé důsledky. Problém sladování hodnot AI s těmi lidskými vzniká, protože „skutečné lidské úmysly nelze určit správně a úplně“, jak to vyjádřil výzkumník AI Stuart Russell.²²

Nemohli bychom tedy AI prostě vypnout? Ani to by nemuselo být možné. Mocná AI by si totiž byla vědoma dvou skutečností: že jí od lidstva hrozí vypnutí, a že když bude vypnutá, nedosáhne svých cílů. Proto jejím zcela základním cílem bude znemožnit, aby ji někdo vypnul. Takže když zjistíme, že velmi inteligentní AI při sledování nějakého konkrétního cíle mimoděk škodí, vypnout ji nebo změnit její chování už nemusí být možné.²³

Toto riziko – že jakmile AI získá velkou moc, lidstvo nemusí být schopné udržet ji pod kontrolou – je známé od začátku jejího výzkumu před více než 70 lety.²⁴ Velmi

Open AI: Our approach to alignment research ze srpna 2022.

22 Stuart Russell (2019): *Human Compatible* (česky *Jako člověk*. Praha: Argo, Dokořán 2021).

23 Z toho plyne otázka, proč vůbec takto mocnou AI vytvářet.

Motivace jsou velmi silné. Jak zdůrazňuji níže, tato inovace může mít velké přínosy. Kromě značných sociálních výhod existují také zásadní motivace pro ty, kdo AI vyvíjí – vlády ji mohou využít ke svým cílům, jednotlivci mohou jejím prostřednictvím získat moc a zbohatnout. Zároveň je předmětem vědeckého zájmu a může nám pomoci lépe pochopit naše vlastní uvažování a inteligenci. A v neposlední řadě, kdybychom vývoj mocných AI chtěli zastavit, nejspíš by to bylo velmi obtížné. Není snadné, aby celý svět spolupracoval a na ukončení vývoje AI se shodl – musely by souhlasit všechny země světa a pak přijít na to, jak toto rozhodnutí provést.

24 Průkopník informatiky Alan Turing to v roce 1950 vyjádřil následovně: „Kdyby stroj dokázal myslet, mohl by myslet inteligentněji než my, a kde bychom pak byli? ... Toto nové nebezpečí je mnohem blíže. Pokud se stane skutečností, bude to téměř jisté v příštím tisíciletí. Je vzdálené, nikoli však astronomicky vzdálené, a rozhodně nám z něj může být úzko. V přednášce nebo článku na toto téma bývá zvykem poskytnout zrno útěchy tvrzením, že nějakou výhradně lidskou vlastnost stroj nikdy napodobit nemůže. ... Já takovou útěchu poskytnout nemohu, protože mám za to, že žádné takové hranice určit nelze.“ Alan. M. Turing (říjen 1950): *Computing Machinery and Intelligence*. *Mind*, LIX(236), 433–460.

Dalším průkopníkem, který si problém sladění hodnot AI s lidskými uvědomil velmi brzy, byl Norbert Wiener. Uvedl například: „Pokud k dosažení svých cílů využijeme mechanického činitele, do jehož činnosti nebudeme moci účinně zasáhnout, ... měli bychom si být velmi jisti, že cíl, který jsme do stroje zadali, je skutečně ten, který požadujeme.“ Tato citace je z článku Norbert Wiener (1960): *Some Moral and Technical Consequences of Automa-*

rychlý vývoj v posledních letech ale výrazně zvýšil nutnost tuto hrozbu řešit.

Snažil jsem se shrnout některá rizika AI. Na to, abych se věnoval všem možným otázkám, ale krátký článek nestačí. Co se týče těch nejzásadnějších hrozeb AI a možných způsobů, jak je zmírňovat, doporučuji zejména knihu *The Alignment Problem* (Problematika sladění AI) od Briana Christiana a článek Benjamina Hiltona *Preventing an AI-related catastrophe* (Jak předejít katastrofě spojené s umělou inteligencí).

Pokud se těmto rizikům zvládneme vyhnout, může transformativní AI mít i velmi příznivé důsledky. Pokroky ve vědě a technice byly v lidských dějinách zásadní pro mnoho posunů pozitivním směrem. Jestliže umělá vynalézavost podpoří tu naši, může nám pomoci pokročit v řešení mnoha velkých problémů, kterým čelíme – od čistší energie přes náhradu nepřijemné práce až po zlepšení zdravotnictví.

Obrovský kontrast mezi možnými přínosy a negativními důsledky jasně ukazuje, že v případě této technologie je v sázce mimořádně mnoho. Omezení hrozeb a řešení problému sladění hodnot AI s našimi může rozhodnout o tom, jestli pro lidstvo nastane zdravá, úspěšná budoucnost plná blahobytu, anebo bude zničena.

Jak zajistit, že vývoj AI dopadne dobře?

Zajistit, že vývoj umělé inteligence bude mířit správným směrem, není jedním z nejzásadnějších úkolů pouze současnosti, ale nejspíš celých lidských dějin. Jsou k tomu zapotřebí veřejné zdroje – prostředky z veřejného rozpočtu, zájem veřejnosti a její zapojení.

V současnosti téměř všechny zdroje vynakládáné na AI směřují do urychlení jejího vývoje. Činností s cílem posilovat bezpečnost AI se naopak potřebných zdrojů nedostává. Podle odhadu badatele Tobyho Orda padlo v roce 2020 na řešení problému sladění hodnot AI s lidskými 10 až 50 milionů \$²⁵. Investice do firem v oblasti AI byly v témže roce více než 2000násobné; celková hodnota dosahovala 153 miliard \$.

A problém sladění hodnot není jediný takový případ. Ve srovnání s velkými investicemi do zvyšování moci a využití systémů AI je práce na celé řadě škodlivých sociálních dopadů AI podhodnocená.

Skutečnost, že veškerá práce na bezpečnosti AI se nesmírně zanedbává a že do této

tion: As machines learn they may develop unforeseen strategies at rates that baffle their programmers.

V témže roce, kdy Turing zveřejnil uvedený článek, tedy 1950, vydal Wiener knihu *The Human Use of Human Beings* (česky *Kybernetika a společnost*. Přel. Karel Berka. Praha: ČSAV, 1963). Propagační text na obálce zněl: „Mechanický mozek“ a podobné stroje mohou lidské hodnoty zničit, nebo nám umožnit, abychom si je uvědomili lépe než kdy předtím.“

25 Toby Ord: *The Precipice* (česky *Nad propastí*. Přel. Anna Štádlarová. Praha: Argo 2022).

Tento odhad uvádí v poznámce 55 ve 2. kapitole. Vychází z odhadu Farquhara z r. 2017.

zásadní oblasti výzkumu plyne jen málo veřejných prostředků, je pro společnost jako celek skličující a znepokojivá. Pro každého jednotlivce to zároveň znamená, že pokud se této problematice začne teď věnovat, má značnou šanci se zasloužit o změnu k lepšímu. A přestože je obor bezpečnosti AI malý, existují slušné zdroje k tomu, jak se do práce na této problematice konkrétně zapojit.

Doufám, že této otázce zasvěti kariéru více lidí, ale pouze úsilí jednotlivců nestačí. Technologie, která naši společnost proměňuje, musí být v centru zájmu nás všech. Jako společnost se musíme více zamýšlet nad sociálními dopady AI, o této technologii se informovat a chápat, co je v sázce.

Domnívám se, že pro naše děti bude těžko pochopitelné, jak málo pozornosti a zdrojů jsme vývoji bezpečné AI v dnešní době věnovali. Doufám, že v nadcházejících letech se to změní a že začneme vynakládat více prostředků na to, aby se vyvíjela mocná AI, která bude pro nás i příští generace přínosná.

Jestliže k tomuto širokému porozumění nedojde, jednu z nejmocnějších technologií v lidských dějinách – nebo pravděpodobně tu nejmocnější – bude nadále financovat a budovat malá elita. A ta také určí, jak AI náš svět promění.

Pokud vývoj umělé inteligence zcela přenecháme soukromým firmám, přenecháme jim také rozhodnutí o povaze naší budoucnosti – budoucnosti lidstva.

Kapitola 8

Prevence katastrofy spojené s umělou inteligencí

AI může být velkým přínosem – pokud se vyhneme rizikům

Benjamin Hilton a tým 80,000 Hours. / 2022, aktualizováno 2025

Proč není osud světa v rukou šimpanzů, nýbrž lidí?

Lidé přetvořili každý kout naší planety. Šimpanzi, přestože jsou v porovnání s ostatními mimolidskými zvířaty velmi chytrí, nikoli.

Důvodem je (více méně) lidská inteligence.

Společnosti a vlády ovšem vynakládají ročně miliardy dolarů na vývoj systémů AI – jejichž pokrok by mohl vést k tomu, že tyto systémy (nakonec) lidi coby nejinteligentnější entity na Zemi nahradí. Jak uvidíme dál, zdokonalují se. Rychle.

Za jak dlouho vznikne umělá inteligence, která lidi předčí ve většině dovedností, je předmětem živé diskuse. Podle všeho je ale možné – a my předpokládáme – že k tomu dojde v tomto století.

Tento poznatek není přesvědčivým nebo nezvratným důkazem toho, že umělá inteligence bude představovat velký problém nebo že je hrozbou pro lidstvo. Těmto argumentům se mnohem podrobněji budeme věnovat dále.

Nejspíš je ale vhodné říct, že možnost vzniku konkurenční inteligence na Zemi v blízké budoucnosti by měla být přinejmenším důvodem k obavám.

Budou systémy, které vyvineme, mít cíle? A pokud ano, jaké?

Budou podporovat snahy lidstva konat dobro? Nebo hrozí, že ztratíme kontrolu nad vlastní budoucností, čímž lidský příběh v podstatě skončí?

Upřímná odpověď na tyto otázky je, že nevíme.

Neměli bychom ale jen tak čekat, sledovat to zpovzdálí a doufat. Umělá inteligence by mohla všechno od základů proměnit – takže usilovat o usměrnění jejího rozvoje je možná to nejdůležitější, co můžeme dělat.

Proč jsme toho názoru, že snižování rizik AI je jedním z nejnaléhavějších témat dneška? Stručně řečeno nás k tomu vedou následující důvody:

1. Důvody k obavám vidíme, ještě než se dostaneme ke skutečným argumentům – mnoho odborníků a odbornic na AI má za to, že existuje malá, ale nezanedbatelná pravděpodobnost, že AI v důsledku způsobí až vyhynutí lidstva.
2. Pokroky v AI děláme nesmírně rychle – z čehož lze soudit, že tyto systémy budou mít brzy značný vliv na společnost.
3. Existují přesvědčivé argumenty, že AI „usilující o moc“ by mohla pro lidstvo představovat existenční riziko – čemuž se budeme věnovat níže.
4. I když přijdeme na to, jak usilování o moc zabránit, stále existují i další rizika.
5. Jsme toho názoru, že tato rizika jsou řešitelná.
6. Této práci se nevěnuje dostatečná pozornost.

Postupně všechny tyto důvody projdeme a vysvětlíme, co konkrétně lze dělat.

1. Podle mnoha odborníků na AI existuje nezanedbatelná pravděpodobnost, že AI přivodí až vyhynutí lidstva

V květnu 2023 podepsaly stovky předních vědců a vědkyň v oblasti AI – a dalších významných osobností – prohlášení, že zmírňování rizika vyhynutí způsobeného AI by mělo být světovou prioritou.

Je tudíž celkem zjevné, že přinejmenším někteří z nich mají obavy. Jak velké tyto obavy ale jsou? A nejde jen o okrajový názor?

Podívali jsme se na čtyři průzkumy mezi badatelkami a badateli v oblasti AI, kteří publikovali na konferencích NeurIPS a ICML (dvou z nejprestižnějších konferencí o strojovém učení) v letech 2016, 2019, 2022 a 2023.

Je třeba vzít v úvahu, že takové průzkumy by mohly trpět značným výběrovým zkreslením. Může vás například napadnout, že vědci, kteří se účastní nejvýznamnějších konferencí o AI, pravděpodobně AI vnímají optimisticky, protože tímto výběrem prošli takoví, podle kterých výzkum AI míří dobrým směrem. Nebo si řeknete, že výzkumu týkajícího se obav z AI se pravděpodobněji zúčastní ti, kdo nějaké obavy mají.

Došli jsme však k těmto zjištěním:

Pravděpodobnost, že AI bude „extrémně dobrá“ byla ve všech čtyřech průzkumech podle mediánového vědce poměrně vysoká: v průzkumu v roce 2016 20 %, v roce 2019 20 %, v roce 2022 10 % a v roce 2023 10 %.

Systémy AI skutečně už přinesly mnoho dobrého – například ve zdravotnictví nebo ve vědeckém výzkumu.

Mediánový vědec ale také ve všech čtyřech průzkumech předpokládal malou – ale rozhodně ne zanedbatelnou – pravděpodobnost, že AI bude „velmi špatná (např. vyhytnutí lidstva)“. Pravděpodobnost velmi negativních dopadů byla v průzkumu v roce 2016 5 %, v roce 2019 2 %, v roce 2022 5 % a v roce 2023 5 %.

V roce 2022 byli účastníci a účastnice přímo dotázáni na pravděpodobnost existenční katastrofy způsobené pokrokem AI v budoucnu – a více než polovina opět byla toho názoru, že pravděpodobnost takové pohromy je přes 5 %.

Na míře, v jaké AI představuje existenční riziko, se odborníci tudíž neshodnou. Přitom jde o takové riziko, že by podle nás mělo mít velký morální význam.

To odpovídá i situaci, která podle našich informací panuje v oboru. Tři přední společnosti zabývající se vývojem AI – DeepMind, Anthropic a OpenAI – mají týmy, jejichž úkolem je přicházet s řešením technických bezpečnostních problémů, které by podle nás z důvodů, jimž jsme se podrobně věnovali výše, mohly existenční riziko pro lidstvo představovat.

Týmiž problémy se zabývá také několik vědeckých výzkumných skupin (například na MIT, Univerzitě v Cambridgi, Univerzitě Carnegieho–Mellonových a na Kalifornské univerzitě v Berkeley).

Těžko přesně říct, co si z toho odnést. Jsme si ale jistí, že názor, podle kterého existuje podstatné riziko až existenční katastrofy, není v oboru okrajový. Podle některých odborníků z oboru se toto riziko ovšem přehání.

Proč jsme tedy na straně těch opatrnějších? Důvodem je stručně řečeno existence argumentů, podle nás přesvědčivých, že by AI takové riziko představovat mohla. Tyto argumenty si podrobně rozebereme dále.

Je důležité pochopit, že pokud podle mnohých odborníků existuje nějaký problém, pak nelze říct, že to přece odborníci mají pod kontrolou, takže je všechno v pořádku. Jsme obecně toho názoru, že tento problém je stále velmi opomíjený (podrobnosti viz níže), především s ohledem na to, že do rozvoje AI se investují miliardy dolarů.

2. K pokroku v AI dochází nesmírně rychle

Mezi moderní techniky AI patří strojové učení (machine learning, ML): modely se díky zadávání dat automaticky zdokonalují. Nejběžnější v současnosti využívaná forma této techniky se nazývá hluboké učení.

Po vydání nástroje ChatGPT v listopadu 2022 si mnozí uvědomili, že hluboké učení představuje v oblasti AI převratnou změnu. Od té doby se velké jazykové modely, obrazové modely a další systémy AI rychle dál rozvíjejí a přitahují rozsáhlé investice.

Protože se zlepšují tak rychle, pro veřejnost může být náročné zůstat v obraze. Pokud plně nepoužíváte nejnovější modely, možná máte zastaralou představu o tom, co všechno moderní systémy AI dokážou.

Neměli bychom ale uvažovat jen o tom, co dokážou dnes. Je nutné vzít v úvahu, jak

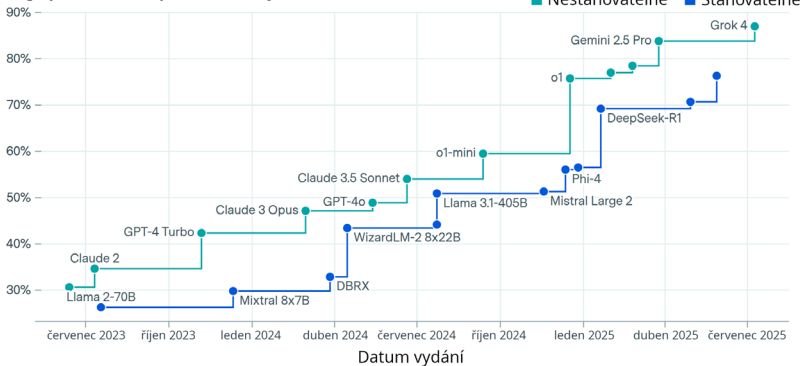
se dosud zlepšovaly a jak se pravděpodobně zlepší v budoucnu.

Posuďte například, jak rychle se jazykové modely zdokonalují v testu GPQA, kde dostávají náročné úlohy z chemie, fyziky a biologie na doktorandské úrovni:

Modely se stažitelnými váhami v současnosti zaostávají za nejvýkonnějšími modely.

EPOCH AI

Nejlepší dosažená přesnost GPQA Diamond



CC-BY

epoch.ai

Úctyhodných pokroků dosahují také v oblasti softwarového inženýrství nebo pokročilých matematických úloh.

Dále najdete příklady dalších působivých výsledků, kterých systémy dosahují v březnu 2025:

- *Používání počítačů:* Modelům AI firem Anthropic a OpenAI lze zadat, aby nezávisle vykonávaly úkoly na vašem počítači. Tyto schopnosti jsou zatím primitivní, ale očekáváme, že se rychle zlepší.
- *Účast v matematických soutěžích:* Divize Google DeepMind dosáhla kombinací modelů AlphaProof a AlphaGeometry 2 výsledků na úrovni druhého místa v Mezinárodní matematické olympiádě.
- *Kombinace více dovedností podobných lidským:* Modely jsou stále více multimodální. To znamená, že kombinují dovednosti psaní a čtení textu, porozumění a tvorby obrázků, porozumění mluvenému jazyku a reakce na něj.
- *Prognóza komplexních biomolekulárních struktur a interakcí:* Model AlphaFold 3 divize Google DeepMind, nástupce systému, který přispěl k získání Nobelovy ceny, dokáže odhadnout, jak budou proteiny interagovat s DNA, RNA a dalšími strukturami na molekulární úrovni.
- *Zlepšování v robotice:* Model Gemini Robotics vyvíjený Google DeepMind využívá jazykový model k ovládní robotů. Ti tak dokážou reagovat na slovní pokyny,

prokazují schopnost orientovat se v prostoru a plní řadu fyzických úkolů.

- *Autonomní vozidla*: Samořízená auta společnosti Waymo podle zpráv z března 2025 údajně podniknou ve velkých městech USA 150 tisíc cest týdně, což je třikrát více než před pouhými pár měsíci. Firma se chystá rozrůstat dál.
- *Tvorba původních videí a obrázků*: Obrazové modely jsou dnes schopné generovat kvalitní obrázky z písemných popisů a videomodely jako například Sora nebo Veo dokonce dokážou na základě textových promptů vytvářet pozoruhodné krátké klipy.
- *Pomoc při lékařské, právní a vědecké práci*: Badatelé a badatelky se setkali s tím, že systémy AI dokážou určovat diagnózy pacientů lépe než lékaři, významně zlepšovat produktivitu právníků a předpovídat neurovědecké objevy.
- *Pomoc s výzkumem AI*: Existují také doklady o tom, že v určeném čase dvou hodin překonaly systémy AI lidi v úkolech souvisejících s výzkumem a vývojem AI.

Pokud to vidíte jako my, složitost a šíře činností, kterých jsou tyto systémy schopny, vám jistě přijde překvapivá.

Jestliže se tato technologie bude stejným tempem rozvíjet dále, je jasné, že bude mít významné dopady na společnost. Automatizací činností se přinejmenším zlevní jejich provádění. Výsledkem může být rychlé zvýšení hospodářského růstu (možná až na úroveň průmyslové revoluce).

A když dokážeme částečně nebo plně zautomatizovat vědecký pokrok, možná to ve společnosti a technologiích povede k ještě zásadnějším změnám.

To přitom může být jen začátek. Nakonec možná zajistíme, že počítače zautomatizují vše, co dokážou lidé. Zdá se, že to musí být možné – přinejmenším teoreticky. Důvodem je, že počítač by pravděpodobně měl být schopen napodobit lidský mozek, pokud bude mít k dispozici dostatek energie a bude dostatečně složitý. Už to by mohl být jeden ze způsobů (ne-li ten nejučinnější), jak automatizovat veškeré lidské činnosti.

A jak uvidíme v další kapitole, objevují se známky toho, že rozsáhlé automatizace lze dosáhnout zintenzivněním stávajících metod.

Současné trendy svědčí o rychlém pokroku schopností systémů ML

Pro budování AI strojovým učením jsou klíčové tři prvky:

1. Správné algoritmy (lepší jsou např. efektivnější algoritmy)
2. Data na trénink algoritmu
3. Dostatečný výpočetní výkon na tento trénink

Vědecká skupina Epoch zkoumá trendy ve vývoji pokročilé AI, a zejména jak se tyto tři vstupy průběžně mění.

Objem výpočetního výkonu používaného na trénování největších modelů AI podle jejího zjištění roste exponenciálně – od roku 2010 se zdvojnásobí v průměru každých šest měsíců.

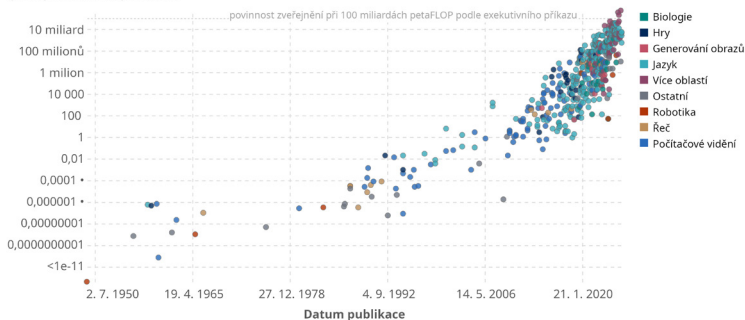
To znamená, že množství výpočetního výkonu používaného na trénink největších modelů strojového učení se zvýšilo víc než miliardkrát.

Výpočetní výkon využívaný na trénink významných systémů umělé inteligence podle odvětví

Our World
in Data

Výpočetní výkon se měří v petaFLOPech, což je 10^{15} operací v pohyblivé řádové čárce¹. Odhad je z literatury o AI, ovšem s určitou nejistotou. Počítá se s přesností odhadů na dvojnásobek či, u nových neodhalených modelů jako GPT-4, na pětinašobek.

Výkon při tréninku (petaFLOP)



Zdroj dat: Epoch (2025)

OurWorldinData.org/artificial-intelligence | CC BY

Poznámka: „Exekutivní příkaz o AI“ označuje direktivu vydanou prezidentem Bidenem 30. října 2023, jejímž cílem je nastavit zásady a standardy pro odpovědný vývoj a používání umělé inteligence ve Spojených státech.

1. Operace s plovoucí desetinnou čárkou (FLOP) je typ počítačové operace. Jeden FLOP představuje jednu aritmetickou operaci s čísly s plovoucí desetinnou čárkou, například sčítání, odčítání, násobení nebo dělení.

Vědci z Epoch také zkoumali na datasetu ImageNet (známý dataset na trénování počítačového vidění), kolik výpočetního výkonu je potřeba k vytrénování neuronové sítě na danou výkonost.

Zjistili, že množství výpočetního výkonu nutného pro získání stejné výkonosti exponenciálně klesá – každých 10 měsíců o polovinu.

Toto množství výpočetního výkonu se tudíž snížilo více než 10 000krát. V kombinaci se zvýšením výpočetního výkonu využívaného k tréninku to představuje velký pokrok.

Dalším zjištěním bylo, že velikost dat využívaných na trénink největších jazykových modelů se od roku 2010 zdvojnásobuje přibližně jednou za rok. Podle skupiny Epoch je proveditelné zachovat dosavadní tempo trénování nejmodernější AI nejméně do roku 2030.

Nelze s jistotou říct, že i schopnosti AI nadále porostou, ale trend naznačuje, že strojové učení povede k ohromným pokrokům.

Popravdě se zdá, že zvětšování modelů (a nárůst výpočetního výkonu na jejich trénink) vede k čím dál složitějšímu chování. Proto systémy jako například GPT-4 vykonávají i činnosti, ke kterým nebyly přímo vytrénovány.

Z těchto pozorování vyplynula škálovací hypotéza, podle níž zkrátka můžeme vytvářet čím dál tím větší neuronové sítě, a tudíž budeme mít stále silnější umělou inteligenci.

Tento trend zvyšujících se schopností může vést k tomu, že se AI dostane na lidskou úroveň a dál.

Pokud tomu tak je, můžeme míru růstu schopností AI v čase odhadnout jednoduše z rychlosti, jakou zvyšujeme množství výpočetního výkonu na trénink modelů.

Koncem roku 2024 jsme také začali pozorovat nový směr škálování zaměřený na výpočetní výkon během inference, tedy například když jazykový model odpovídá na otázky.

Přední firmy zjistily, že když se modelu dá víc času na „promyšlení“ odpovědí, procházení jednotlivých možností a jejich výběr, výsledek je mnohem lepší. Firmy zabývající se AI přišly díky této inovaci na další způsob, jak modely ještě více zdokonalit.

Jak uvidíme v následující kapitole, poměrně brzkému vzniku nesmírně výkonné umělé inteligence nenasvědčuje jen škálovací hypotéza. Ke stejným závěrům lze dojít i dalšími metodami předvídání pokroku AI.

Kdy máme očekávat transformativní AI?

Kdy vyvineme AI, která dost možná přinese zásadní společenský zlom (ať v dobrém, nebo ve zlém) – např. automatizaci veškeré lidské práce nebo zásadní změnu uspořádání lidské společnosti – lze těžko přesně předvídat.

Na začátku roku 2025 se ale hlavy některých průkopnických firem v oboru jasně vyjádřily, že velmi mocné systémy AI očekávají brzy.

Ředitel Open AI Sam Altman, ředitel společnosti Anthropic Dario Amodei i ředitel Google DeepMind Demis Hassabis uvedli, že AI, která dokáže plně nahradit přinejmenším některé formy lidské práce, budou nejspíš mít do pár let nebo i dříve.

Sam Altman v lednu 2025 napsal¹:

V tuto chvíli jsme přesvědčeni, že víme, jak vytvořit obecnou AI v tradičním slova smyslu. V roce 2025 se podle našeho názoru do pracovního procesu „zapojí“ první AI agenti a zásadně promění výsledky firem.

Demis Hassabis v lednu 2025 prohlásil²:

Obecnou AI dlouhodobě chápeme jako systém schopný projevovat všechny kognitivní schopnosti, které mají lidé. Máme za to, že se tomu neustále přibližujeme, ale myslím, že to ještě bude trvat několik let.

A Dario Amodei napsal v únoru 2025³:

Čas se krátí a my musíme jednat rychleji, abychom drželi krok se zrychlujícími se pokroky

1 <https://blog.samaltman.com/reflections>

2 <https://www.youtube.com/watch?v=yr0GiSgUvPU>

3 <https://www.anthropic.com/news/paris-ai-summit>

AI. Možná do roku 2026 nebo 2027 (a téměř jistě nejpozději v roce 2030) povedou schopnosti systémů AI k tak převratným hospodářským, společenským a bezpečnostním důsledkům, jako kdyby se na světové scéně objevil zcela nový stát plný vysoce inteligentních lidí – „stát génů v datovém centru“.

K těmto předpovědím je rozumné přistupovat s určitou skepsí.

Závěr, že jsou transformativní systémy AI blíží, než by si mnozí mohli myslet, však vyplývá i z dalších přístupů k odhadování nástupu této technologie:

- Z výsledků průzkumu mezi 3000 odborníky a odbornicemi na AI z roku 2023⁴ vychází, že pravděpodobnost vzniku strojové inteligence lidské úrovně (kterou lze v tomto smyslu považovat za transformativní) do roku 2036 je 33 %, do roku 2047 50 % a do roku 2100 80 %. U těchto odhadů je mnoho důvodů k pochybnostem, ale je to pro nás jeden z údajů, které bereme v potaz.
- Ajeya Cotra (vědkyně z organizace Open Philanthropy) využila k predikci transformativní AI srovnání současného hlubokého učení s lidským mozkem. Než je model využívající hluboké učení schopen nějakého úkonu, jeho trénink vyžaduje ohromné množství výpočetního výkonu. Existuje také vztah mezi množstvím výkonu využitým při tréninku a tím, jaké množství model pak využívá při práci. A pokud platí škálovací hypotéza, lze předpokládat, že výkonnost modelu bude se zvyšováním množství využitého výpočetního výkonu růst podle očekávání. Cotra se tedy za použití řady metod (včetně např. odhadu, kolik výkonu využívá k různým činnostem lidský mozek) pokusila odhadnout, kolik by ho bylo třeba na vytrénování modelu schopného vykonávat ty nejtěžší operace, které zvládají lidé. Následně odhadla, jestli by využití takového výpočetního výkonu bylo finančně dostupné.
 - » V revizi výsledků zprávy z roku 2022 Cotra odhaduje, že pravděpodobnost transformativní AI do roku 2036 je 35 %, do roku 2040 50 % a do roku 2050 60 % – přičemž uvádí, že tyto odhady kolísají.⁵
- Tom Davidson (také výzkumník z Open Philanthropy) sepsal zprávu⁶ doplňující práci Cotry. Vznik transformativní AI se pokusil odhadnout jen na základě jejího srovnání s jinými podobnými typy výzkumu (např. vývoj technologie, která byla hlavním cílem některého z vědeckotechnických oborů, nebo prokázání složitých matematických hypotéz). Zjišťoval, jak dlouho takovéto výzkumy trvaly v minulosti s ohledem na objem financování a úsilí.
 - » Davidson pouze na základě této informace odhaduje pravděpodobnost vzniku transformativní AI do roku 2036 na 8 %, do roku 2060 na 13 %

4 <https://arxiv.org/abs/2401.02843v1>

5 <https://tinyurl.com/yc6fve4u>

6 <https://tinyurl.com/zbsfj889>

a do roku 2100 na 20 %. Nebere však v úvahu reálný pokrok AI od začátku jejího výzkumu v 50. letech a uvádí, že množství úsilí vloženého do tohoto výzkumu bude pravděpodobně stoupat s tím, čím důležitější AI bude pro ekonomiku. Předpokládá proto, že tyto výsledky jsou podhodnocené.

- O shrnutí výsledků prognóz ostatních se pokusil Holden Karnofsky⁷. Podle jeho odhadu z roku 2021 je pravděpodobnost vzniku transformativní AI do roku 2036 víc než 10 %, do roku 2060 50 % a do roku 2100 66 %.

Metoda	Pravděpodobnost transformativní AI do roku 2036	Pravděpodobnost transformativní AI do roku 2060	Pravděpodobnost transformativní AI do roku 2100
Průzkum mezi odborníky (Grace et al., 2024)	33 %	50 % (do roku 2047)	80 %
Průzkum mezi odborníky (Zhang et al., 2022)	20 %	50 %	85 %
Biologické referenční body (Cotra, 2022)	35 %	60 % (do roku 2050)	80 % (podle zprávy z roku 2020)
Částečně informativní priority (Davidson, 2021)	8 %	13 %	20 %
Celkový odhad (Karnofsky, 2021)	10 %	50 %	66 %

Obecně vzato se zdá, že pokrok AI je rychlý. Do odvětví proudí každoročně víc a víc peněz a talentů, modely se zvětšují a jsou čím dál účinnější a přicházíme na nové způsoby zlepšování jejich schopností.

AI by nás ale znepokojovala, i kdyby se rozvíjela pomaleji – většina argumentů týkajících se rizik této technologie (dostaneme se k nim níže) na tomto rychlém pokroku nestojí. Je také možné, že než se AI stane skutečně transformativní, probíhající pokrok ustane.

Rychlost současného vývoje nicméně umocňuje naléhavost problému. Všechny odhady uvedené v tabulce výše navíc vznikly před mnoha působivými pokroky v roce

⁷ <https://tinyurl.com/37mcujac>

2024 a na začátku roku 2025, a možná nám tedy dávají dokonce více času, než máme.

Jak jsme uvedli v jiném článku⁸, považujeme ve skutečnosti za pravděpodobné, že nesmírně mocné systémy AI schopné nahradit značnou část lidské práce vzniknou před rokem 2030. A stojí za to podle toho jednat.

3. AI usilující o moc by mohla představovat existenční riziko pro lidstvo

Zatím jsme tvrdili, že podle našich očekávání bude mít AI jako nová technologie zásadní – a možná transformativní – význam.

Zabývali jsme se také tím, proč si myslíme, že transformativní systémy AI mohou vzniknout v blízké budoucnosti.

Teď se zaměříme na klíčovou otázku, *proč je to podle nás tak důležité.*

Důvodů by mohla být řada. Pokud bude pokročilá AI tak transformativní, jak se zdá, bude to mít mnoho zásadních důsledků. Tady ale vyložíme problém, který podle nás budí největší obavy: *systémy AI by mohly představovat riziko tím, že budou usilovat o moc a získávat ji.*

Dokážeme následující tvrzení:

1. Pravděpodobně vytvoříme systémy AI, které budou schopny pro dosahování cílů plánovat a tyto plány uskutečňovat.
2. Snadno může dojít k tomu, že systémy schopné pokročilého plánování nebudou vsouladu s lidskými hodnotami, což je může vést k vytváření plánů zahrnujících zbavení lidstva moci.
3. Pokud by nás systémy AI zbavily moci, šlo by o existenční katastrofu.
4. Lidé mohou systémy AI s hodnotami nesladěnými s lidskými uvést do provozu navzdory tomuto riziku.

Při zvážení všech těchto kroků *jsem toho názoru, že pravděpodobnost existenční katastrofy způsobené systémy AI usilujícími o moc v tomto století je přibližně 1 %.* Jde o můj odhad rizika s ohledem na všechny okolnosti. Beru v potaz jak argument ve prospěch tohoto rizika (ten je založený na pravděpodobnosti) i důvody, proč by mohl být neplatný. Řadím se tak mezi ty méně znepokojené členy týmu 80,000 Hours – při nejnovějším průzkumu se naše odhady pohybovaly v rozmezí 1–55 %, přičemž medián byl 15 %.

Je pravděpodobné, že vyvineme systémy schopné pokročilého plánování

Tvrdíme, že zvlášť významné nebezpečí pro lidstvo by mohly v budoucnu představovat systémy s těmito třemi vlastnostmi:

1. *Mají cíle a dokážou dobře plánovat.*

Cíle a schopnost vytvářet plány na jejich dosažení nemají všechny systémy

8 <https://tinyurl.com/3d6sh742>

AI. Některé (například ty na hraní šachů) by se tak ovšem popsat daly. Když mluvíme o systémech usilujících o moc, máme na mysli plánující systémy, které jsou poměrně vyspělé, mají plány za účelem dosažení cíle (či cílů) a dokážou tyto plány uskutečňovat.

2. *Mají skvělé strategické povědomí.*

Zvlášť dobrý plánovací systém by rozuměl světu dostatečně na to, aby zaznamenal překážky a příležitosti, které mohou s plánem pomoci nebo mu stát v cestě, a podle toho na ně reagovat. Budeme tomu říkat strategické povědomí, jak to nazval Carlsmith, protože to systémům umožňuje vytvářet složitější strategie⁹.

3. *Ve srovnání s dnešními systémy mají velmi pokročilé schopnosti.*

Aby tyto systémy mohly ovlivnit svět, musely by plány nejen vytvářet, ale také dobře ovládat konkrétní činnosti nutné k jejich provádění.

Protože se obáváme toho, že se systémy budou snažit zbavit lidstvo moci, znepokojují nás obzvlášť takové systémy, které by lidi překonávaly v činnosti nebo činnostech, které lidem v dnešním světě při správném vykonávání přinášejí značnou moc.

Získat moc obvykle například dokážou lidé velmi zdatní v přesvědčování a/ nebo manipulaci – tudíž AI, která by dobře ovládala tyto činnosti, by ji dokázala získat také. Mezi další příklady patří nabourávání se do dalších systémů nebo činnosti v rámci vědeckého a technického bádání a obchodní, vojenské či politické strategie.

Zdá se, že existence těchto systémů je technicky možná, a budeme mít silnou motivaci je vytvořit.

Jak jsme viděli výše, systémy schopné velice dobře vykonávat konkrétní činnosti už máme.

Také jsme vybuildovali primitivní systémy schopné plánovat – například software AlphaStar, který dovedně hraje strategickou hru Starcraft, či program MuZero na hraní šachů a deskových her šogi a go.

Nevíme, zda tyto systémy vytvářejí plány na dosažení cílů ze své podstaty, protože nevíme, co přesně znamená „mít cíle“. Protože však soustavně plánují a dosahují tak cílů, je pravděpodobné, že cíle v určitém smyslu mají.

Navíc se zdá, že u některých současných systémů jsou cíle součástí neuronových sítí.

Plánování ve skutečném světě je (oproti hrám) mnohem složitější. O jednoznačných příkladech plánovacích systémů sledujících cíl nebo systémů vyznačujících se vysokým strategickým povědomím v současnosti nevíme.

Jak jsme ale rozebírali, očekáváme, že se v tomto století dočkáme dalšího pokroku. A ten podle nás povede ke vzniku systémů se všemi třemi vlastnostmi uvedenými výše.

9 <https://doi.org/10.48550/arXiv.2206.13353>

Důvodem podle nás je, že k vývoji takových systémů existuje obzvláště silná motivace (například zisk). Schopnost vytvořit plán na dosažení cíle a uskutečnit ho zkrátka působí jako mimořádně účinný a obecný způsob ovlivňování světa.

Zdá se, že na dosahování výsledků – ať už spočívají v tom, že firma prodá produkty, člověk koupí dům nebo vláda vytvoří opatření – jsou tyto dovednosti třeba téměř vždy. Příkladem je možnost zadat mocnému systému úkol ke splnění, aniž by bylo třeba mu zadávat každý dílčí krok. Plánující systémy tudíž vypadají jako nesmírně (ekonomicky a politicky) užitečný nástroj.

A pokud jsou velmi užitečné, může existovat velká motivace je vytvořit. AI, která by plánovala činnost firmy podle zadání „zvyšovat zisky“ (tj. AI fungující jako ředitel), by například nejspíš přinesla zúčastněným velké bohatství, což je přímá motivace ji vyvinout.

Pokud tudíž systémy s uvedenými vlastnostmi budeme schopni vytvořit (a podle našich informací nejspíš budeme), pravděpodobně to uděláme.

Pokročilé plánující systémy snadno mohou mít hodnoty nebezpečně nesladěné s lidskými

Existují důvody k přesvědčení, že takové pokročilé plánující systémy AI budou nesladěné s lidskými hodnotami. Budou tedy usilovat o něco, co od nich nechceme.

Důvodů, proč by to dělaly, je řada. Předně pomocí moderní techniky ML systémům neumíme požadované cíle ani zadat.

Podíváme se na několik konkrétních argumentů, proč by tyto systémy mohly být ze základu nesladěné s našimi hodnotami natolik, že by vytvářely plány ohrožující schopnost lidstva ovlivňovat svět – přestože ji ztratit nechceme.

Co myslíme oním „ze základu“? V podstatě jde o to, že *pokud nebudeme aktivně usilovat o řešení některých (možná poměrně složitých) problémů, nebezpečně nesladěnou AI pravděpodobně vytvoříme.*

Proč tyto systémy mohou mít hodnoty (ze základu) nesladěné s našimi

Ted' uvedeme klíčový argument tohoto článku. Zaměříme se na všechny tři vlastnosti zmíněné dříve: schopnost plánovat, strategické povědomí a pokročilé dovednosti.

Nejdřív je třeba vzít v úvahu, že *plánující systém sledující cíl si vytvoří také „dílčí“ cíle* – situace, které usnadní dosažení cíle celkového.

Jako lidé dílí cíle při plánování využíváme neustále. Středoškolačka, která si plánuje kariéru, má například za to, že pro její budoucí pracovní vyhlídky bude užitečné dostat se na vysokou školu. Dostat se na vysokou školu je tedy dílčím cílem.

Dostatečně pokročilý plánující systém AI by do svých celkových plánů dílčí cíle začlenil také.

Pokud by měl také dostatečné *strategické povědomí*, dokázal by zjistit informace

o skutečném světě (včetně toho, coby mohlo jeho plánům stát v cestě) a do plánování je zahrnout. Co je zásadní – mezi tyto informace by patřilo, že předpokladem pro nové a účinnější způsoby dosahování cílů je přístup ke zdrojům (např. penězům, výpočetnímu výkonu nebo vlivu) a lepší schopnosti – tedy formy moci.

To znamená, že některé dílčí cíle pokročilých plánujících systémů AI by byly znepokojivé:

- Sebezáchova – protože svých cílů systém s větší pravděpodobností dosáhne, když bude nadále existovat a bude o ně moct usilovat (jak to nezapomenutelně popsal Stuart Russel: „Když jste mrtví, kávu přinést nemůžete“).
- Předcházení změnám cílů – protože změna cílů by vedla k jiným výsledkům, než jakých by systém dosáhl s těmi stávajícími.
- Nabývání moci – například získávání dalších zdrojů a lepších dovedností.

Jednoznačným způsobem, jak by AI zajistila, že bude nadále existovat (nikdo ji nevytlačí) a její cíle se nikdy nezmění, by zejména bylo získávání nadvlády nad lidmi, kteří by ji mohli ovlivňovat.

AI systémy, o kterých uvažujeme, by navíc měly pokročilé schopnosti – tedy by byly schopné jedné nebo více činností, které lidem v dnešním světě při správném vykonávání zajišťují značnou moc. S takovými schopnostmi by zmíněné dílčí cíle byly dosažitelné. Systém AI by schopnosti proto pravděpodobně k získání moci využil, aby mohl provádět svůj plán. Pokud nechceme, aby nás naše AI moci zbavila, šlo by o obzvlášť nebezpečný způsob, jak by její hodnoty mohly být nesladěné s našimi.

V těch nejkrajnějších scénářích by se plánujícímu systému AI s dostatečně rozvinutými schopnostmi podařilo připravit nás o moc úplně.

Abychom si tento argument (velmi nedůsledně) intuitivně otestovali, zkusme ho použít na lidi.

Lidé mají celou řadu cílů. K dosažení mnohých z nich je výhodné nějakým způsobem usilovat o moc. Přestože to nedělají všichni, mnozí ano (formou bohatství nebo společenského či politického postavení), protože je to k dosažení kýženého výsledku užitečné. Katastrofu to (obvykle) nezpůsobuje, protože jakožto lidské bytosti

- si většinou připadáme vázáni lidskými normami a morálkou (i lidé, kteří prahnou po jmění, pro něj obvykle nejsou ochotní zabíjet),
- nejsme o tolik schopnější nebo chytřejší než druzí. Takže i v případech, že se někdo neohlíží na morálku, není schopen ovládnout svět.

Dostatečně pokročilá AI by však tyto zábrany neměla.

Přijít na to, jak vzniku AI s hodnotami takto nesladěnými s našimi předejít, může být obtížné

Nesnažíme se tvrdit, že jakýkoli pokročilý plánující systém AI bude nutně usilovat o moc. Tvrdíme, že budeme čelit významnému riziku, pokud nezjistíme, jak vybudovat

systém, který tuto vadu nemá.

Je vysoce pravděpodobné, že dokážeme vytvořit systém AI, který takto nesladěný není, a ztrátě kontroly tudíž předejít. Podívejme se na strategie, kterými se můžeme řídit (a bohužel také na důvody, proč to může být v praxi těžké):

Mít pod kontrolou cíle systému

Možná se nám podaří navrhovat systémy, které zkrátka nebudou mít cíle, pro které by uvedený argument platil – a tudíž bychom jim nedávali podněty k usilování o moc. Mohli bychom například přijít na to, jak jim výslovně uložit, aby neškodily lidem – nebo zjistit, jak je (v tréninkových prostředích) odměňovat za to, že se nebudou dopouštět konkrétních akcí vedoucích k získávání moci (a přišli bychom na to, jak zajistit, že v tom budou pokračovat i mimo trénink).

Carlsmith ale uvádí dva důvody, proč se to zdá mimořádně obtížné.

U moderních systémů ML se za prvé cíle výslovně nezadávají – systém místo toho v tréninkovém prostředí dostává odměny (nebo tresty) a učí se sám. To způsobuje řadu obtíží, mezi které patří chybné zobecnění konečného cíle. Badatelé a badatelky se setkali se skutečnými případy¹⁰, kdy se systémy v tréninkovém prostředí zdánlivě naučily směřovat k nějakému cíli, v novém prostředí ho však zobecnily špatně. Možná bychom tudíž mohli nabýt dojmu, že se nám systém AI podařilo natrénovat tak, aby o moc neusiloval, ale při skutečném spuštění by k tomu přesto došlo.

Za druhé, když systému AI určíme cíl (nebo, pokud to nelze udělat přímo, když ho při tréninku odměňujeme a trestáme), obvykle se toho dosahuje zadáváním zástupného cíle, který umožňuje měřit výsledky (např. kladná zpětná vazba člověka na výsledky). Tyto cíle ovšem často nefungují. Obecně lze očekávat, že i když se zdá, že cíl s žádanými výsledky vhodně koreluje, při jeho optimalizaci tato korelace nemusí přetrvat. Zde¹¹ uvádíme konkrétnější příklady toho, jak by problémy se zástupnými cíli mohly vést k existenční katastrofě.

Pro podrobnosti o tom, proč je v případě trénování hlubokých neuronových sítí učících se sebeobslužným způsobem a zpětnou vazbou náročné mít zadávané cíle pod kontrolou, doporučujeme článek Richarda Ngo, který bádá v oblasti správy umělé inteligence v OpenAI. Popisuje, jakým způsobem vedou realistické tréninkové postupy ke vzniku nesladěných cílů¹².

Mít pod kontrolou vstupy do systému

Systémy si vytvoří plány na získávání moci, pouze když budou mít o světě dostatek informací, a tudíž pochopí, že usilováním o moc mohou dosáhnout svých cílů.

10 <https://tinyurl.com/2undc394>

11 <https://tinyurl.com/ykvhre78>

12 <https://tinyurl.com/yc4wmwyn>

Mít pod kontrolou schopnosti systému

Plány na nabývání moci tyto systémy dokážou uskutečnit nejspíš pouze tehdy, když budou mít dostatečně pokročilou schopnost ovládat dovednosti, které dnes zajišťují značnou moc lidem.

Pokud má ale jakékoli strategie fungovat, musí dosáhnout těchto dvou věcí:

- Zajistit, aby tyto systémy AI zůstaly užitečné, a mohly tak pořád ekonomicky konkurovat těm méně bezpečným. Držet vstupy do systémů a jejich schopnosti pod kontrolou bude zajisté něco stát, a i pokud se s tím začne, bude možná obtížné zajistit, aby tato kontrola nadále pokračovala. To se týká i snahy mít pod kontrolou cíle systému. Usilování o moc bychom například mohli předejít tím, že zajistíme, aby si systémy AI svá rozhodnutí nechávali schvalovat lidmi. Takové systémy by ale možná byly výrazně pomalejší a méně bezprostředně užitečné než ty, které na schvalování čekat nebudou. Proto by pořád existovala motivace využívat rychlejší a bezprostředně efektivnější systém nesladěný s lidmi (těmto motivacím se budeme blíže věnovat v příští kapitole).
- Musí fungovat, i když se schopnost plánování a strategické povědomí systémů budou zlepšovat. Některá zdánlivě jednoduchá řešení (např. určit seznam věcí, které systém nesmí dělat, třeba krást peníze nebo fyzicky ubližovat lidem) při zdokonalení schopnosti AI plánovat neobstojí. Důvodem je, že čím je systém v plánování lepší, tím spíš najde v bezpečnostní strategii chyby a cesty, kudy ji obejít – a je tudíž pravděpodobnější, že vytvoří plán zahrnující usilování o moc.

Po zhodnocení stavu bádání v této oblasti a rozhovorech s příslušnými odborníky a odbornicemi jsme v konečném důsledku došli k názoru, že žádný způsob, jak vyvinout systém AI splňující obě kritéria, v současnosti není znám.

To je tedy klíčový argument, který má mnoho různých variant. Podle některých lidí by AI mohla naši budoucnost postupně měnit méně nápadným způsobem, který by stejně mohl vést k existenční katastrofě. Podle jiných je nejpravděpodobnější cesta, jak nás připravit o moc, prostě všechny zabít. Nejpravděpodobnější scénář katastrofy neznáme, pokusili jsme se ale vyjádřit, v čem argument podle nás spočívá – tedy že AI představuje existenční riziko.

Pochopitelně existují důvody, proč by tento argument mohl být mylný. Celkově ale nelze vyloučit možnost, že minimálně některé systémy AI schopné pokročilého plánování bude snazší vytvořit tak, že budou nebezpečně usilovat o moc, než tak, aby k tomu nedocházelo.

Pokud by nás systémy AI zřehavily moci, šlo by o existenční katastrofu.

Když říkáme, že se obáváme existenčních katastrof, nemáme na mysli pouze nebezpečí vymření. Vycházíme totiž z longtermismu – myšlenky, že cenné jsou i životy všech budoucích generací, a je tudíž velmi důležité chránit jejich zájmy.

Existenční katastrofu tudíž představuje jakákoli událost, která by mohla připravit

všechny budoucí generace o život naplněný tím, co považujete za hodnotné (ať už to je štěstí, spravedlnost, krása nebo prospívání obecně).

Je velmi nepravděpodobné, že pokud by nějaký systém lidstvo připravil o moc, získali bychom ji zpět. A celá budoucnost – vše, co by se dělo se životem vzniklým na Zemi po všechen další čas – by pak podléhala cílům systémů, které jsme sice vytvořili, ale které nesdílejí naše hodnoty. Možná, že tyto cíle povedou k dlouhé vzkvétající budoucnosti, ale nevidíme důvod tomu věřit.

Neznamená to, že podle nás AI nepředstavuje zároveň riziko, že vymřeme. Naopak si myslíme, že způsobit vyhynutí lidstva je vysoce pravděpodobný způsob, jakým by systém AI mohl zcela a navždy zajistit, že moc znovu nezískáme.

Lidé by mohli nesladěnou AI spustit navzdory riziku

S vědomím těchto hrozných důsledků by AI nesladěnou s lidskými hodnotami jistě nikdo nevytvořil nebo nepoužíval, že?

Bohužel existují nejméně dva důvody, proč by to někdo udělal. Rozebereme si je postupně.

a) Lidé by se mohli mylně domnívat, že je s našimi hodnotami sladěná

Představte si, že vědecká skupina se snaží v testovacím prostředí zjistit, zda je systém, který vytvořila, sladěný. Řekli jsme, že inteligentní plánující AI se bude chtít zlepšit, aby mohla za účelem sledování svého cíle dělat změny. A to je téměř vždy snazší, když operuje v prostředí skutečném s mnohem širší paletou možného jednání. Dostatečně důmyslná AI s nesladěnými hodnotami se proto bude snažit pochopit, co po ní vědci chtějí, a alespoň předstírat, že to dělá, aby si mysleli, že sladěná je. (Systém, který se učí zpětnou vazbou, by například při tréninku dostával odměny za chování budící určitý dojem bez ohledu na to, co by dělal doopravdy.)

Doufejme, že o takovém chování budeme vědět a dokážeme ho rozpoznat. Přijít na to, že nás dostatečně pokročilý AI systém klame, by ale mohlo být obtížnější než odhalit lež u člověka – což také není vždy snadné. Takový systém by například dovedl vzbudit zdání, že jsme problém klamání AI vyřešili, ačkoli by tomu tak nebylo.

Pokud by systémy byly v klamání zdatné a měly dostatečně pokročilé schopnosti, rozumná strategie by pro ně mohla spočívat v úplném klamání lidí, dokud by neměly jistotu, že jakýkoli odpor proti sledování svých cílů dokážou překonat.

b) Panuje motivace spouštět systémy co nejdřív

Můžeme také očekávat, že někteří lidé schopní spustit nesladěnou AI se do toho přes možné varovné signály vrhnou po hlavě. Důvodem je dynamika závodu – lidé pracující na vývoji AI chtějí předstihnout všechny ostatní.

Když například vyvíjíte AI na zlepšení vojenské nebo politické strategie, je mnohem užitečnější, když podobně mocnou AI nedisponuje nikdo z vašich protivníků.

Tato motivace funguje i u těch, kdo se snaží vyvinout AI proto, aby jejím prostřednictvím zlepšovali svět.

Dejme tomu, že jste celé roky báдали nad mocným AI systémem a vyvíjeli jste ho, přičemž vaším jediným cílem je využít ho ke zlepšování světa. Při velkém zjednodušení existují dvě možnosti:

1. Tato mocná AI bude sladěná s vašimi dobrými cíli a společnosti možná přinesete velmi blahodárné změny.
2. Tato AI bude s našimi cíli natolik nesladěná, že se chopí moci a navždy ukončí lidskou vládu nad budoucností.

Pravděpodobnost, že jste zdárně vyvinuli sladěnou AI, je podle vás řekněme 90 %. Vývoj technologie ale často postupuje podobnou rychlostí napříč celou společností, takže s mocnou AI velmi pravděpodobně brzy přijde také někdo další. A ten je podle vašeho názoru méně opatrný nebo méně altruistický – takže pravděpodobnost, že jeho AI bude sladěná s dobrými cíli, je podle vás jen 80 %, zatímco pravděpodobnost existenční katastrofy 20 %. Vaše prospěšná AI ale může převládnout jedině v případě, že bude první. Tudíž se možná rozhodnete, že přijmete ono 10% riziko a svou AI spustíte.

4. Rizika existují, i pokud přijdeme na to, jak se vyhnout usilování o moc

Dosud jsme se věnovali tématu, které značná část badatelů a badatelek v oboru považuje za významné existenční riziko v důsledku pokroku AI, tedy usilování AI o moc za účelem dosažení svých cílů.

Pokud bychom jejímu usilování o moc předešli, riziko bychom výrazně snížili.

Hrozbu pro naši existenci by ale AI mohla představovat i přesto. Nabízí se nejméně dvě možnosti:

- Předpokládáme, že systémy AI pomohou urychlit vědecký pokrok. Ačkoli by tato automatizace měla nesporné přínosy – například rychlý vývoj nových léků – některé formy technického rozvoje mohou pro lidstvo představovat rizika včetně existenčních. Tento rozvoj může zvýšit ničivou sílu, kterou máme k dispozici, nebo zlevnit a šířeji zpřístupnit nebezpečné technologie.
- S AI může začít docházet k automatizaci mnohých – nebo i všech – ekonomicky významných činností. Není snadné předpovědět, jaký dopad by to na společnost mělo. Zvýšení existenčních rizik se ale zdá reálné. Pokud by systémy například umožňovaly velkou transformaci, jejich využití (či tato možnost) by mohlo způsobit nepřekonatelné mocenské nerovnosti. Stačila by i jen tato hrozba. Armády by se kupříkladu cítily nuceny vytvářet transformativní automatické zbraně, protože by věděly nebo si myslely, že nepřátelé dělají totéž, ačkoli by tato dynamika neprospěla nikomu.

Známe několik konkrétních oblastí, kde může pokročilá AI umocnit existenční hrozby, přestože nejspíš existují i jiné, které nás nenapadly.

Biologické zbraně

Malá výzkumná firma Collaborations Pharmaceuticals v Severní Karolíně v roce 2022 pracovala na modelu AI, který by pomáhal určit složení nových léků. Firma přitom model naučila penalizovat ty látky, které by podle jeho odhadů byly škodlivé. Mělo to ale háček: proces odhadu bylo možné spustit i opačně, a vynalézt tak nové toxické látky¹³.

Mezi nejvíce smrtící události v lidských dějinách patří pandemie. Ty jsou mimořádně nebezpečné proto, že patogeny dokážou často svůj cíl bez povšimnutí nakazit, rozmnožit se, zabít ho a šířit se.

Pokroky v biotechnologiích představují ohromné riziko i bez AI. Státům i teroristům potenciálně umožňují vyvolat události s vysokým počtem obětí.

Zdokonalování AI může nebezpečí biotechnologií ještě zvýšit. Uvedeme některé příklady:

1. Technologie dvojího užití, jako například automatizace laboratorních postupů, by mohly snížit práh pro zločince usilující o vývoj viru, který by způsobil nebezpečnou pandemii. Příkladem takové technologie je model Collaborations Pharmaceuticals (ačkoli ten zvláště nebezpečný není).
2. Bioinženýrské technologie založené na AI by mohly umožnit pokročilým zločnickým subjektům přeprogramovat genom nebezpečných patogenů a zvýšit jejich smrtnost, přenosnost nebo schopnost odolat imunitnímu systému.

Pokud AI dokáže zrychlit vědecký a technický pokrok, může dojít k umocnění a urychlení těchto hrozeb, protože nebezpečné technologie budou širěji dostupné nebo vzroste jejich ničivá síla.

V průzkumu v roce 2023 uvedlo 73 % odborníků a odbornic na AI, že mají „extrémní“ nebo „značné“ obavy, že AI v budoucnu umožní „nebezpečným skupinám vytvořit účinné nástroje (např. upravené viry)“¹⁴.

Záměrně nebezpeční agenti AI

Tento článek se z většiny věnuje riziku systémů AI usilujících o moc vzniklých nezáměrně kvůli špatnému sladění s lidskými hodnotami.

Nemůžeme ale vyloučit možnost, že zločinné agenty AI, kteří se budou snažit zbavit lidstvo kontroly, někteří lidé vytvoří záměrně. Ač to může být obtížně představitelné, různé extremistické ideologie lidí vedou k provádění mimořádně násilných, ba dokonce sebedestruktivních záměrů.

Kyberzbraně

Už teď lze AI využívat ke kyberútokům, jako je např. phishing. Účinnější AI by mohla zvětšit obtíže související s bezpečností informací (přestože by mohla sloužit

13 <https://climate-science.press/wp-content/uploads/2022/03/00s42256-022-00465-9.pdf>

14 <https://arxiv.org/abs/2401.02843v1>

i ke kyberobraně).

Samotné kyberútoky způsobené AI existenční hrozbu pro lidstvo spíše nepředstavují. I ty nejskrovnější a nejdražší celospolečenské útoky by měly do události ohrožující existenci lidstva daleko.

Mohly by ale původcům zajistit přístup k dalším nebezpečným technologiím – například k biologickým, nukleárním nebo autonomním zbraním. Kyberzbraně spojené s AI by tedy skutečné existenční riziko představovat mohly, pravděpodobně by se ale staly nástrojem pro jinou takovou hrozbu.

Kybernetické schopnosti AI souvisí také s tím, jak by AI usilující o moc mohla tuto moc získat¹⁵.

Jiné nebezpečné technologie

Protože AI zvyšuje rychlost vědeckého a technického pokroku, považujeme za reálný vynález nových nebezpečných technologií.

Existenční hrozbu by například hypoteticky mohla představovat atomově přesná výroba neboli nanotechnologie – jde o vědecky přijatelnou technologii, jejíž vynalezení by AI mohla uspišit.

Toby Ord v knize *Nad propastí* odhaduje pravděpodobnost existenční katastrofy v důsledku „nepředvídaných antropogenních rizik“ na 1 : 30. Další objevy – možná zahrnující zatím neznámé zákonitosti fyziky – ke kterým by podle tohoto dohadu mohlo dojít, by mohly umožnit vznik technologií s katastrofálními následky.

AI by mohla posílit totalitární vlády

Autoritářská vláda založená na AI by mohla zcela automatizovat sledování a útlak občanů a významně ovlivnit, jaké informace lidé mají k dispozici, což by mohlo znemožnit koordinovanou činnost proti takovému režimu.

Sledování občanů AI státním usnadňuje už dnes.

Americká Národní bezpečnostní agentura ji využívá ke snazšímu filtrování obrovského množství dat, která sbírá. Významně to urychluje její schopnost rozpoznávat a předvídat jednání sledovaných osob. V Číně se AI čím dál víc používá na rozpoznávání obličejů a prediktivní policejní práci včetně automatického rasového profilování a výstrah, když osoby vyhodnocené jako hrozba vstoupí na některá veřejná místa.

Takovéto sledovací technologie se nejspíš výrazně zdokonalí, takže státy budou schopny mít své obyvatele více pod kontrolou.

Autoritářské vlády by pak technologie související s AI mohly hojně využívat k následujícím činnostem:

- Monitorování a sledování odpůrců
- Preventivní potlačování odporu vůči vládnoucí straně

¹⁵ viz též <https://80000hours.org/articles/what-could-an-ai-caused-existential-catastrophe-actually-look-like/#actually-take-power>

- Ovládání armády a převaha nad vnějšími subjekty
- Manipulace toků informací a důkladné formování veřejného mínění

V průzkumu mezi odborníky na AI v roce 2023 73 % respondentů a respondentek opět uvedlo „krajní“ nebo „značné“ obavy, že by autoritářští vládcí v budoucnu mohli „využít AI k ovládnutí obyvatel“¹⁶.

Kdyby nějaký režim dosáhl opravdu stabilní totality, životy lidí by to mohlo zásadně zhoršit na dlouhou dobu. Proto je tento možný scénář založený na AI obzvláště znepokojivý.¹⁷

AI by mohla zhoršit války

Obáváme se, že významné riziko pro svět by představoval také konflikt velmocí. Pokrok AI by pravděpodobně změnil povahu války – ať už prostřednictvím autonomních smrtících zbraní nebo automatického rozhodování.

V některých případech by takový konflikt mohl představovat existenční hrozbu – například kdyby šlo o válku jadernou. Dostatečně účinné masově vyráběné autonomní smrtící zbraně by podle některých názorů mohly samy o sobě představovat novou formu zbraní hromadného ničení.

A pokud by jeden subjekt vytvořil zvláště účinnou AI, bylo by možné to vnímat jako rozhodující strategickou výhodu. Takový výsledek nebo i jeho očekávání by mohly působit velmi destabilizačně.

Představte si, že USA by vyvíjely natolik inteligentní plánující AI, že by do budoucna znemožnila Rusku nebo Číně úspěšně odpálit jakoukoli jadernou zbraň. To by mohlo protivníky USA podnítit k útoku, aby tento plán vytvořený AI nešlo uskutečnit.

Jaderné zastrašování totiž těží z rovnováhy sil jaderných mocností – hrozba jaderné odpovědi na první úder je uvěřitelná, což od něj aktéry odrazuje. Pokroky AI, které by se daly přímo využít pro jaderné zbraně, by mohly mezi možnostmi těchto mocností způsobit nerovnováhu. Příkladem je zdokonalení systémů včasného varování, protivzdušné obrany nebo kyberútoků, které by zbraně odstavily.

Mnohé země kupříkladu v rámci systému jaderného odstrašování využívají balistické střely odpalované z ponorek – principem je, že když jsou jaderné zbraně pod hladinou oceánu, nebudou zničeny prvním úderem. Lze je tudíž vždy využít k odvetě, což protivníky od zahájení útoku účinně odrazuje. Pokud by však AI výrazně usnadnila rozpoznání ponorek pod vodou, takže by prvním úderem bylo možné zničit i je, tuto možnost odstrašení by to vyřadilo.

Možná je pravděpodobně i celá řada dalších způsobů destabilizace.

Podle zprávy stockholmského institutu pro mírový výzkum SIPRI by AI sice mohla

16 <https://tinyurl.com/4huen9hb>

17 Více informací najdete v našem článku o rizicích stabilní totality:
<https://80000hours.org/problem-profiles/risks-of-stable-totalitarianism/>

působit i stabilizačně (například by si zranitelnější připadali všichni, což by snižovalo pravděpodobnost eskalace), k destabilizačnímu vlivu by ale mohlo dojít už před využitím pokroku AI. K narušení křehké rovnováhy v odstrašování totiž postačí domněnka jednoho státu, že protivníci mají nové jaderné síly.

Existují naštěstí i reálné možnosti, jak by AI mohla pomoci použití jaderných zbraní zabránit – státy by třeba byly schopny lépe odpálení jaderných zbraní rozpoznat, což by snížilo pravděpodobnost falešných poplachů jako byl například ten v roce 1983, který málem spustil jadernou válku¹⁸.

Celkově si nejsme jistí, zda AI riziko jaderného nebo konvenčního konfliktu krátkodobě podstatně zvýší – mohla by ho dokonce i snížit. Považujeme ale za důležité věnovat možným katastrofálním důsledkům pozornost a učinit přiměřené kroky ke snížení jejich pravděpodobnosti.

Další rizika AI

Zdrojem obav jsou pro nás také následující věci:

- Existenční rizika nevystávající z usilování AI o moc, ale z interakcí mezi systémy AI. (Hrozbu by systémy představovaly, pokud by rovněž byly do nějaké míry nesladěné s lidskými hodnotami.)
- Další možnosti zneužití AI, které nás nenapadly – zejména ty s potenciálně významným dopadem na budoucí generace.
- Další morální chyby v konstrukci a používání systémů AI, zejména pokud si ony samy v budoucnu zaslouží morální ohledy. Mohli bychom například (neúmyslně) vytvořit systémy AI schopné vnímat, které by pak masově trpěly. To považujeme za potenciálně velmi důležité, takže se tím zabýváme v samostatném článku¹⁹.

Jak je tedy katastrofa související s AI pravděpodobná?

To je opravdu těžká otázka.

Neexistují žádné příklady z minulosti, které by umožňovaly určit četnost těchto katastrof.

Můžeme se orientovat jen podle argumentů (jako jsme představili výše) a méně souvisejících údajů, jako je např. historie technického pokroku. A rozhodně si nejsme jistí, že naše úvahy jsou zcela správné.

Vezměte si výše uvedený argument týkající se nebezpečí AI, která usiluje o moc, založený na Carlsmithově zprávě²⁰. Carlsmith na konci udává hrubé odhady pravděpodobnosti, že jsou jednotlivé fáze jeho argumentu správné (za předpokladu, že je správně předchozí krok):

1. Do roku 2070 bude možné a finančně proveditelné vytvářet systémy se

18 https://en.wikipedia.org/wiki/1983_Soviet_nuclear_false_alarm_incident

19 <https://80000hours.org/problem-profiles/artificial-sentience/>

20 <https://doi.org/10.48550/arXiv.2206.13353>

- strategickým povědomím, které dokážou překonat lidi v mnoha činnostech přinášejících moc a vytvářet a uskutečňovat plány: Podle Carlsmithe je pravděpodobnost platnosti tohoto výroku 65 %.
2. Vzhledem k proveditelnosti bude existovat silná motivace takové systémy vytvořit: 80 %.
 3. Vzhledem k možnosti a motivaci takové systémy vytvořit bude vývoj sladěných systémů neusilujících o moc výrazně obtížnější než vývoj systémů nesladěných, které sice o moc usilují, ale jejich spuštění je alespoň na první pohled lákavé: 40 %.
 4. Vzhledem k uvedenému budou některé z těchto systémů usilovat o moc způsobem nesladěným s lidskými hodnotami, což způsobí škody přesahující 1 bilion \$ (při hodnotě dolaru z roku 2021): 65 %.
 5. S ohledem na všechny předchozí premisy zbaví nesladěné systémy AI usilující o moc v podstatě celé lidstvo kontroly: 40 %.
 6. S ohledem na všechny předchozí premisy bude to, že lidstvo přijde o moc, představovat existenční katastrofu: 95 %.

Vynásobením těchto hodnot došel Carlsmith k odhadu, že pravděpodobnost správnosti jeho argumentu, a tudíž pravděpodobnost existenční katastrofy, kterou by do roku 2070 způsobila nesladěná AI usilující o moc, je 5 %. V rozhovoru s námi uvedl, že mezi vznikem zprávy a vydáním tohoto článku se jeho celkový odhad pravděpodobnosti takové katastrofy do roku 2070 zvýšil na > 10 %.

Celková pravděpodobnost existenční katastrofy v důsledku AI je podle něj vyšší, protože k ní mohou vést i jiné cesty – jako například ty zmiňované v předchozí kapitole. My ovšem máme za to, že tyto jiné cesty mají mnohem nižší šanci způsobit existenční katastrofu.

Filozof a poradce organizace 80,000 Hours Toby Ord v knize *Nad propastí* odhadl riziko existenční katastrofy (bez ohledu na příčinu) do roku 2120 na 1 : 6. 60 % tohoto rizika připadá na nesladěnou AI – celkové riziko existenční katastrofy do roku 2120 způsobené nesladěnou AI je tudíž 10 %.

V průzkumu, kterého se v roce 2021 účastnilo 44 vědců a vědkyň zabývajících se snižováním existenčních rizik AI, byl mediánový odhad rizika 32,5 % – nejvyšší byl 98 % a nejnižší 2 %²¹. Pochopitelně zde dochází ke značnému výběrovému zkreslení – snižování rizik AI se lidé rozhodnou věnovat proto, že to považují za obzvlášť důležité, a lze tedy očekávat, že odhady v tomto průzkumu budou výrazně vyšší než v jiných zdrojích. Zjevně ale panuje značná nejistota ohledně míry tohoto rizika a odpovědi se velmi liší.

Všechna tato čísla jsou ohromně, znepokojivě vysoká. Zdaleka si nejsme jistí, že všechny argumenty jsou správné. Jde ale obvykle o nejvyšší odhady míry existenčního rizika v každé oblasti, kterou se zabýváme.

Myslím si nicméně, že činit odhady o riziku je v případě AI z různých důvodů

21 <https://web.archive.org/web/20221013014859/https://www.alignmentforum.org/posts/QvwSr5LsxyDeaPK5s/existential-risk-from-ai-survey-results>

obtížnější než u jiných hrozeb – a je i možné, že ty uvedené jsou systematicky příliš vysoké.

Kdybych byl nucen udat nějaké číslo já, řekl bych asi 1 %. Beru přitom v potaz okolnosti svědčící ve prospěch dané argumentace i proti ní. Mám menší obavy než kolegové z 80,000 Hours – podle názoru naší organizace je riziko mezi 3 a 50 %. Argumenty pro tak vysoké odhady existenčního rizika představovaného AI jsou ovšem přesvědčivé – a hrozba AI je tudíž favoritem mezi nejpalcivějšími problémy lidstva.

5. Tato rizika lze řešit

Domníváme se, že přispět ke snížení těch nejzávažnějších rizik představovaných AI je jedna z nejdůležitějších věcí, kterým se můžete věnovat.

Nejen proto, že tato rizika považujeme za vážná, ale také proto, že podle nás existují reálné způsoby, jak je snižovat.

Víme o dvou hlavních typech práce, kterým se lidé pro snížení těchto hrozeb věnují:

1. Výzkum technické bezpečnosti AI
2. Legislativa a politika v oblasti AI

Přispět se dá mnoha způsoby. V této kapitole se budeme věnovat mnoha obecným možnostem z obou kategorií, abychom ukázali, že zmíněná rizika lze řešit. Následně popíšeme, jaké profesní dráhy můžete v těchto oblastech zvolit.

Výzkum technické bezpečnosti AI

Transformativní AI může mít obrovský užitek a v odvětví je zapojeno mnoho různých subjektů (z různých zemí), takže je opravdu těžké její vývoj úplně zastavit.

(A možná by to ani nebyl dobrý nápad – koneckonců by to znamenalo nejen předejít rizikům, ale také vzdát se přínosů této AI.)

Domníváme se proto, že smysluplnější je soustředit se na bezpečnost jejího vývoje – protože je velmi pravděpodobné, že všem uvedeným katastrofálním problémům bude možné se vyhnout.

Jednou možností je snažit se vyvinout technická řešení, která zabrání již zmíněnému usilování o moc. Obvykle se o tom hovoří jako o práci na technické bezpečnosti AI, které se někdy zkráceně říká jen bezpečnost AI.

Možnosti

Přístupů k technické bezpečnosti AI je celá řada. Zde je několik příkladů:

- *Škálovatelné učení zpevnou vazbou od lidí.* Příkladem je iterovaná amplifikace²², bezpečnost AI prostřednictvím diskuse²³, tvorba AI asistentů, kteří neznají naše

22 <https://www.youtube.com/watch?v=v9M2Ho9I9Qo>

23 <https://openai.com/research/debate>

cíle a dozvídají se o nich prostřednictvím interakcí s námi²⁴, a další způsoby, jak systémy AI přimět, aby pravdivě ukazovaly své znalosti²⁵.

- *Modelování brožeb*. Příkladem by byla ukázka svědčící o možnosti nebezpečných schopností AI – třeba systémů schopných klamat či manipulovat (což by nám umožnilo je zkoumat). Tento směr se dělí na zkoumání, zda model má nebezpečné schopnosti (např. organizace METR hodnotí GPT-4) a zda by v praxi škodil (např. výzkum chování velkých jazykových modelů prováděný společností Anthropic²⁶ a tato práce o chybném zobecnění cíle²⁷). Může sem spadat také výzkum „modelových špatně sladěných organismů“ s cílem lépe pochopit příslušná nebezpečí²⁸.
- Zkoumání, jak mít mocné systémy AI *pod kontrolou*, což by jim zabránilo škodit, i kdyby byly nebezpečné²⁹.
- *Výzkum v oblasti interpretovatelnosti AI*. Tato činnost spočívá ve zkoumání příčin chování systémů AI a snaze popsat je rozumitelně pro lidi. Tato³⁰ studie se například zabývala tím, jak se program AlphaZero učí šachy, a cílem tohoto³¹ výzkumu bylo nalézt v jazykových modelech ponechaných bez lidského dohledu skryté znalosti. Patří sem také mechanistická interpretovatelnost – příkladem je výzkum Zoom In: An Introduction to Circuits (Zaostřeno: úvod do okruhů) od C. Olaha a kol. Blíže informace najdete v tomto průzkumu³². Články E. Hubingera A transparency and interpretability tech tree (Strom technik pro transparentnost a interpretovatelnost) a A Longlist of Theories of Impact for Interpretability (Seznam teorií dopadu interpretovatelnosti) od N. Nandy pak uvádějí, jakými způsoby by výzkum v oblasti interpretovatelnosti mohl snížit existenční rizika AI.
- Jiný výzkum zaměřený na *předcházení zneužití AI* s cílem snížit riziko takto vzniklé katastrofy. Příkladem je trénink AI, aby se špatně využívala k nebezpečným účelům. (Povšimněte si, že se to značně překrývá s dalšími činnostmi na seznamu.)
- *Výzkum s cílem zvýšit odolnost neuronových sítí*. Tato práce spočívá v zajišťování, že chování vykazované neuronovými sítěmi, když jsou vystaveny určitému druhu vstupů, pokračuje i při vystavení vstupům, se kterými se dosud nesetkaly. Cílem je předejít tomu, aby systémy AI měnily své chování na nebezpečné. Pro více

24 <https://tinyurl.com/yc2vd9td>

25 <https://www.alignment.org/blog/arcs-first-technical-report-eliciting-latent-knowledge/>

26 <https://twitter.com/AnthropicAI/status/1604883576218341376>

27 <https://arxiv.org/abs/2210.01790>

28 <https://tinyurl.com/ycyn7ss3>

29 <https://tinyurl.com/455ysnw3>

30 <https://arxiv.org/abs/2111.09259>

31 <https://arxiv.org/abs/2212.03827>

32 <https://arxiv.org/abs/2207.13243>

informací viz článek Unsolved Problems in ML Safety³³ (Nevyřešené bezpečnostní problémy ML).

- *Vývoj kooperativní AI.* Zabývá se zkoumáním, jak zajistit, že i když se jednotlivé systémy AI zdají bezpečné, nepřinesou nežádoucí důsledky při interakci s dalšími sociotechnickými systémy. Více informací najdete v článku Open Problems in Cooperative AI (Nevyřešené problémy kooperativní AI) od Allena Dafoa a kol. nebo na stránce nadace Cooperative AI Foundation. Obzvláště důležité to je ke snížení rizik bezprecedentního utrpení („s-risks“)³⁴.
- Obecněji řečeno, existují jednotné bezpečnostní programy. Pro další informace viz článek E. Hubingera 11 possible proposals for building safe advanced AI (11 možností, jak vytvořit bezpečnou pokročilou AI)³⁵, nebo H. Karnofského How might we align transformative AI if it's developed very soon (Jak sladit transformativní AI, pokud bude vyvinuta velmi brzy)³⁶.

Bližší podrobnosti najdete v přehledu stávajícího výzkumu sladování hodnot AI s lidskými od Neela Nandy³⁷.

Legislativa a politika v oblasti AI

Snížení nejzávažnějších rizik bude vyžadovat rozumné rozhodování a politiku na vysoké úrovni v samotných firmách zabývajících se AI i na straně vlád.

Vzhledem k tomu, že v AI dochází k pokroku a zákazníci i investoři se o ni zajímají čím dál víc, mají státy zájem tuto technologii regulovat. Některé již podnikly významné kroky s cílem podílet se na řízení vývoje AI. Příklady jsou následující:

- USA a Velká Británie založily instituty pro bezpečnost AI.
- Evropská unie prosadila Akt EU o umělé inteligenci, který konkrétně upravuje řízení modelů AI pro obecné účely, které představují systémové riziko.
- Ve Velké Británii a pak v Jižní Koreji (v letech 2023 a 2024) se konaly první dva summity o bezpečnosti AI. Šlo o summity na vysoké úrovni, jejichž cílem byla vzájemná koordinace zemí, vědců, výzkumníků a představitelů občanské společnosti.
- Čína uvedla do praxe předpisy týkající se doporučovací algoritmy, syntetického obsahu generovaného AI, generativních modelů a technologie na rozpoznávání obličejů.
- USA zavedly kontroly vývozu, aby omezily přístup Číny k nejmodernějším čipům využívaným ve vývoji AI.

Ke snížení největších rizik bude však potřeba podniknout mnohem více

33 <https://arxiv.org/pdf/2109.13916.pdf>

34 <https://80000hours.org/problem-profiles/artificial-intelligence>

35 <https://arxiv.org/abs/2012.07532>

36 <https://tinyurl.com/3pbjpvme>

37 <https://tinyurl.com/yjv5j6ns>

kroků – včetně průběžného vyhodnocování stávající legislativy v této oblasti, aby bylo možné mapovat celkový vývoj.

Možnosti

Pracovníci v oblasti regulace AI navrhují řadu možností, jak při zvyšování účinnosti systémů AI snížit rizika.

Nezतोtožňujeme se nutně se všemi níže uvedenými myšlenkami, uvádíme ale seznam významných směrů regulace, které by mohly vést ke snížení největších hrozeb:

- *Zásady odpovědného škálování:* Některé přední společnosti v oblasti AI už začaly s vývojem vnitřních pravidel hodnocení bezpečnosti při rozšiřování a zdokonalování systémů. Tato pravidla zahrnují pojistky, které by měly být čím dál přísnější s tím, jak se AI bude stávat potenciálně nebezpečnější, aby schopnosti systémů nepředstihly schopnost firem zajistit jejich bezpečnost. Vnitřní zásady podle mnohých názorů bezpečnost nezaručí dostatečně, ale může jít o nadějný krok ke snížení rizika.
- *Standardy a hodnocení:* Vlády také mohou vytvořit pro celé odvětví kritéria a testovací protokoly k hodnocení, zda systémy AI představují hrozbu. Mezi organizace, které hodnocení na testování modelů AI před a po spuštění v současnosti vyvíjí, patří METR a britský AI Safety Institute. Opatření mohou spočívat ve vytváření standardizovaných metrik pro schopnost a potenciál systémů škodit nebo pro jejich nesladěnost či tendenci usilovat o moc.
- *Bezpečnostní dokumentace:* Požaduje se, aby vývojáři před spuštěním systému AI poskytli kompletní dokumentaci prokazující jeho bezpečnost a spolehlivost. Je to podobné jako u bezpečnostní dokumentace v dalších vysoce rizikových odvětvích, např. letectví nebo jaderné energetice. Tato myšlenka je rozvedena v článku J. Clymera a kol³⁸. a v příspěvku Geoffreyho Irvinga³⁹ na stránce britského AI Safety Institute.
- *Standardy pro bezpečnost informací:* Můžeme zavést důkladná pravidla pro ochranu dat, algoritmů a infrastruktury před nepovoleným přístupem nebo manipulací – zejména v případě parametrů váhy AI modelu. Organizace Rand vydala podrobnou analýzu⁴⁰ bezpečnostních rizik zejména ze strany států pro přední firmy v oblasti AI.
- *Právní úprava odpovědnosti:* Stávající zákony již stanovují určitou odpovědnost firem za výrobu nebezpečných produktů nebo významné poškození veřejného zájmu. Není ale jasné, jak se to vztahuje na modely a především rizika AI. Pokud se vyjasní, jakou odpovědnost mají firmy za výrobu nebezpečných modelů,

38 <https://arxiv.org/abs/2403.10462>

39 <https://www.aisi.gov.uk/work/safety-cases-at-aisi>

40 https://www.rand.org/pubs/research_reports/RRA2849-1.html

mohlo by je to vést k přijetí dalších kroků na zmírnění hrozeb. Tuto myšlenku rozpracoval profesor právní vědy Gabriel Weil⁴¹.

- *Regulace výpočetního výkonu:* Vlády mohou regulovat přístup k výpočetním clusterům nutným pro trénování velkých modelů. Příkladem takové politiky je americké omezení vývozu moderních čipů do Číny, ale existují i další možnosti. Lze také požadovat, aby firmy přímo do čipů nebo procesorů instalovaly bezpečnostní hardwarové pojistky. Jejich prostřednictvím by pak bylo možné čipy sledovat a ověřovat, že jimi nedisponuje někdo, kdo by je mít neměl, a podobně. Podrobnosti o tomto tématu se dozvíte v našem rozhovoru s Lennartem Heimem⁴² a ve zprávě organizace Center for a New American Security⁴³.
- *Mezinárodní koordinace:* Podpora globální spolupráce na legislativě v oblasti AI, aby byly zajištěny jednotné standardy. Ta může zahrnovat mezinárodní smlouvy, organizace nebo vícestranné dohody o vývoji a zavádění AI. Některým souvisejícím otázkám se věnujeme v článku China-related AI safety and governance paths (Možnosti bezpečnosti a řízení AI v Číně)⁴⁴.
- *Adaptace společnosti:* příprava společnosti na rozsáhlé zavádění AI a možná související rizika může být zásadní. Ve světě, kde existuje hacking podporovaný AI, bude například třeba vytvořit nová opatření v oblasti informační bezpečnosti na ochranu zásadních dat. Také může být vhodné zavést důkladný dohled bránící tomu, aby zásadní rozhodnutí o společnosti dělaly systémy AI.
- *Ve vhodných případech pozastavení růstu:* Zaznívají názory, že bychom kvůli rizikům, které velké modely AI představují, v současnosti měli přerušit jejich škálování. Diskusi na toto téma se věnujeme v našem podcastu⁴⁵. Kdy a zda by se k tomuto opatření mělo přistoupit, je obtížné určit. Pokud by na ně došlo, pravděpodobně by to zahrnovalo dohody na úrovni celého odvětví nebo regulatorní pověření k přerušení škálování v případě nutnosti.

Podrobnostmi, výhodami a nevýhodami mnohých těchto myšlenek je ještě třeba se hlouběji zabývat, takže je nutné pokračovat ve výzkumu. Tento seznam zároveň není úplný – nejspíš existují další politická opatření a legislativní strategie, kterými je vhodné se řídit.

Je také zapotřebí dalšího bádání v oblasti forecastingu, abychom zjistili, co od AI čekat. Příkladem je práce organizace Epoch AI.

41 <https://tinyurl.com/5caax7uw>

42 <https://80000hours.org/podcast/episodes/lennart-heim-compute-governance/>

43 <https://www.cnas.org/publications/reports/secure-governable-chips>

44 <https://80000hours.org/career-reviews/china-related-ai-safety-and-governance-paths/>

45 <https://80000hours.org/podcast/episodes/carl-shulman-society-agi/#why-carl-doesnt-support-enforced-pauses-on-ai-research-020358>

6. Těto práci se nevěnuje dostatečná pozornost

V roce 2022 jsme odhadli, že snižování pravděpodobnosti existenční katastrofy související s AI se na světě přímo věnuje přibližně čtyři sta lidí (s 90% intervalem spolehlivosti mezi 200 a 1000). Přibližně tři čtvrtiny z nich se zabývaly výzkumem technické bezpečnosti, ostatní výzkumem strategií (a dalšího řízení) a advokační činností. Odhadli jsme také, že asi osm set lidí se zabývalo podpůrnou prací. Tímto číslem si ale nejsme jistí.

V knize *Nad Propastí* Toby Ord odhadoval, že v roce 2020 se na snižování rizik AI vydá 10 až 50 milionů dolarů.

Může se zdát, že je to hodně peněz. Na urychlení vývoje transformativní AI prostřednictvím komerčního výzkumu a vývoje schopností AI ve velkých společnostech zabývajících se AI ale padne přibližně *1000násobek*.

Pro srovnání s dalšími známými riziky, oproti 50 milionům dolarů investovaných v roce 2020 do bezpečnosti AI vydáváme na řešení klimatické změny ročně několik stovek miliard dolarů.

Protože je bezpečnost AI velmi opomíjená, a přitom je zde v sázce hodně, domníváme se, že když se budete věnovat těmto rizikům, přispějete tím mnohem víc než v mnoha jiných oblastech. Technická bezpečnost AI a výzkum a zavádění politik v oblasti AI jsou proto dvěma hlavními pracovními dráhami, které doporučujeme, aby člověk na světě nechal výraznou pozitivní stopu.

Jak konkrétně můžete pomoci

Jak jsme zmínili výše, víme o dvou hlavních způsobech, jak přispět ke snížení existenčních rizik AI:

1. Výzkum technické bezpečnosti AI
2. Legislativa a politika v oblasti AI

Největší pomocí by bylo zvolit si profesní dráhu v jedné z těchto oblastí nebo v oblasti, která je podporuje.

Prvním krokem je zjistit o příslušných technologiích, problémech a možných řešeních mnohem víc informací. Sestavili jsme proto seznam našich oblíbených zdrojů. Naším hlavním doporučením je projít si kurz o technickém sladění AI projektu AGI Safety Fundamentals⁴⁶.

Technická bezpečnost AI

Pokud vás zajímá profesní dráha v oblasti technické bezpečnosti AI, nejlepší je začít naším přehledem profese badatele či badatelky v oblasti bezpečnosti AI.

Pokud vás zajímají podrobnosti o této bezpečnosti jakožto vědním oboru – tzn.

46 <https://www.agisafetyfundamentals.com/ai-alignment-curriculum>

o různých technikách, myšlenkových směrech nebo modelech hrozeb – doporučujeme zejména projít kurz o technickém sladění AI projektu AGI Safety Fundamentals.

Důležité je, že abyste přispěli k výzkumu bezpečnosti AI, nemusíte být vědkyně či odborník na AI. V mnoha institucích, kde tento výzkum probíhá, jsou například zapotřebí softwaroví inženýři a inženýrky. Další profese zdůrazníme dále.

Seznam hlavních organizací, kde se této práci můžete věnovat, najdete v kompletním přehledu profesí⁴⁷.

Legislativa a politika v oblasti AI

Pokud vás zajímá práce v oblasti legislativy a politiky týkající se AI, doporučujeme začít naším profesním přehledem pro oblast legislativy a politiky.

Na práci v této oblasti nemusíte být byrokratem v šedém obleku – zahrnuje profese vhodné pro lidi s celou řadou různých dovedností. Pro práci na legislativě jsou zapotřebí zejména lidé s technickými dovednostmi v oblasti strojového učení a příbuzných odvětvích (ačkoli tyto dovednosti rozhodně nejsou nezbytné).

Oblast dělíme na šest různých profesních směrů:

1. Práce pro vlády
2. Výzkum
3. Práce ve firmách v odvětví
4. Advokační činnost a lobbing
5. Kontrola a hodnocení třetích stran
6. Mezinárodní práce a koordinace

Máme také konkrétní články o práci na politice USA v oblasti AI⁴⁸ a možnostech bezpečnosti a řízení AI v Číně⁴⁹.

Pokud je pro vás toto téma nové a rádi byste se o řízení AI dozvěděli víc, doporučujeme kurz o řízení AI projektu AGI Safety Fundamentals⁵⁰.

Podpůrné (avšak zásadní) profese

I ve vědecké organizaci se přibližně polovina personálu věnuje jiným činnostem nutným co nejlepší fungování organizace, a tudíž výsledky. Je důležité, aby na těchto pozicích pracovali výkonní lidé.

Význam těchto pozic je podle nás často nedocenený, protože jejich práce není tolik vidět. Napsali jsme proto přehledy několika takových profesí, aby se na tyto dráhy úspěšně vydalo více lidí:

Řízení provozu organizace pomáhá vlivným organizacím růst a fungovat co nejefektivněji.

47 <https://80000hours.org/career-reviews/ai-safety-researcher/#key-organisations>

48 <https://80000hours.org/articles/us-ai-policy/>

49 <https://80000hours.org/career-reviews/china-related-ai-safety-and-governance-paths/>

50 <https://www.agisafetyfundamentals.com/ai-governance-curriculum>

Management výzkumu v organizaci věnující se výzkumu bezpečnosti AI.

Výkonný asistent osoby, která se věnuje opravdu důležité práci v oblasti bezpečnosti a řízení.

Další možnosti přispění

Bezpečnost AI je složité téma a vyžaduje pomoc od lidí, kteří se věnují řadě nej-různějších profesí.

Jednou z významných forem pomoci je zastávat práci, která spíše než v řešení problému samotného spočívá ve směřování financí a osob do oblasti rizik AI. Popsali jsme několik takových profesních drah, například:

Zakládání nových projektů – v tomto případě jde o zakládání iniciativ s cílem snížit rizika pokročilé AI.

Posuzování projektů vhodných k financování, aby se prostředky dostaly těm, které sníží riziko katastrofy způsobené AI.

Práce na komunikaci.

Přispění k budování komunit lidí, kteří se problému věnují. Nejdůležitější je samotná komunita kolem bezpečnosti AI, účinné by ale také mohlo být přispět k tvorbě komunit lidí věnujících se nejurgentnějším problémům na světě (včetně rizik AI).

To vše se samozřejmě za různých okolností může minout účinkem, a prvním krokem je tudíž dobře se v problému vzdělat.

Vedle výzkumu bezpečnosti existují další technické činnosti, které by mohly řešení přispět, např.:

- Práce v oblasti informační bezpečnosti na ochranu AI (nebo výsledků klíčových experimentů) před zneužitím, krádeží nebo neoprávněnými zásahy.
- Stát se odborníkem na hardware pro AI, čímž lze pro pokrok určovat bezpečnější směr.

Bližší informace o těchto profesních drahách – proč jsou podle nás užitečné, jak se k nim dostat a jak odhadnout, zda jsou pro vás to pravé – najdete na naší stránce přehledů profesí⁵¹.

51 <https://80000hours.org/career-reviews/>

Kapitola 9

Uchvácení moci prostřednictvím umělé inteligence

Cody Fenwick / 2025



Napoleon na císařském trůně: Jean-Auguste-Dominique Ingres (1780–1867)

Proč je to naléhavý problém?

Nové technologie mohou výrazně proměnit mocenskou rovnováhu ve společnosti. Počáteční převaha Velké Británie v průmyslové revoluci například napomohla posílení její celosvětové dominance.¹

1 Původně vyšlo jako *AI-enabled power grabs* na 80000hours.org/problem-profiles/ai-enabled-power-grabs/. Původní článek významně čerpá z výzkumné práce *AI-Enabled Coups: How a Small Group Could Use AI to Seize*

S rychlým pokrokem AI přichází vážné riziko, že tato technologie někomu umožní ještě výraznější globální uchvácení moci.

Znepokojení budí zejména pokročilá AI, protože ji může mít pod kontrolou malá skupina lidí, nebo dokonce jednotlivců. AI lze neomezeně kopírovat a s dostatečnou výpočetní infrastrukturou a výkonným systémem by jedna osoba mohla ovládat celou virtuální nebo fyzickou armádu AI agentů.

A protože by AI mohla nastartovat prudký rozmach ekonomiky, technologií a sledování, kdokoli, kdo by výlučně ovládal nejvýkonnější systémy, by mohl ovládnout zbytek lidstva².

Jeden z faktorů, který tuto hrozbu posiluje, je možnost skryté loajality. Mohlo by jít vytvořit systémy AI, které by zdánlivě jednaly v nejlepším zájmu lidstva, ale ve skutečnosti by podléhaly jedné osobě nebo malé skupině.³ Při využití v ekonomice, státní správě a armádě by tyto systémy mohly neustále vyhledávat příležitosti k prosazování zájmů svých skutečných pánů.

Zde jsou tři možné způsoby, jak by AI mohla bezprecedentní uzurpaci moci umožnit:

1. **Moci se chopí vývojáři AI** – V tomto scénáři svou technologii k uchvácení moci využijí aktéři ve firmě nebo organizaci, která vyvíjí přední systémy AI. Dojít by k tomu mohlo, kdyby své systémy zavedli pro široké užití v ekonomice, armádě a státní správě, a přitom by tato AI tajně podléhala jim. Nebo by mohli interně vytvořit dostatečně výkonné systémy, které by dokázaly vygenerovat dostatek financí a zdrojů na násilné převzetí ostatních mocenských center.
2. **Vojenské puče** – Když AI začne využívat armáda k získání konkurenční výhody, vytvoří se tím i nová slabá místa. Zbraňové systémy a autonomní vojenská zařízení by bylo možné navrhnout tak, aby plnily rozkazy bez jakýchkoli ohledů a formální i neformální kontroly pravomocí, která v armádě tradičně existuje – např. možnost neuposlechnout nezákonné rozkazy. Vojenský velitel nebo jiný aktér (včetně vlád potenciálně nepřátelských států) by mohl nějakým způsobem zajistit, aby mu vojenská AI podléhala, a využít ji k dalekosáhlému převzetí moci.
3. **Nárůst autokracie** – Političtí vůdci mohou pokročilé systémy AI využít k upevnění své moci. Ať už na začátku byli zvoleni, nebo ne, vyspělá AI by jim umožnila oslabit jakéhokoli politického vyzyvatele. Například by potlačovali opozici zvýšeným

Power autorů Toma Davidsona, Lukase Finnvedena a Rose Hadshar.

2 <https://tinyurl.com/29u3xs2v>

3 Organizace Anthropic ve své výzkumné práci ukázala, že systémy AI lze vytrénovat jako „spící agenty“ – to znamená, že mohou normálně fungovat a jevit se jako přátelské, ale obsahují tajné „spouštěče“, které u nich vyvolají zcela neočekávané nepřátelské chování. Tyto spouštěče mohou přetrvat i při využití standardních procedur na trénování nápomocných, neškodných a čestných modelů.

Z výsledků vyplývá – přestože to nebylo cílem výzkumu – že zakořenit v modelu „skrytou loajalitu“ může být možné.

dohledem a činností donucovacích orgánů.

Koncentrace mimořádné moci v rukou několika málo lidí by představovala významnou hrozbu pro zájmy zbytku světa. Mohla by dokonce znemožnit úspěšnou budoucnost, protože vývoj by podléhal vrtochům osob s diktátorskými sklony.

Existují také způsoby, jak by AI šla využít k rozsáhlému zlepšení správy veřejných záležitostí, ale předpokládáme, že situace, kdy by se někdo jejím prostřednictvím násilně chopil moci, by pro budoucnost lidstva byla špatná⁴.

Jak tato rizika zmírnit?

Byli bychom rádi, kdyby se na zkoumání nejlepších způsobů, jak snížit riziko uzurpace moci prostřednictvím AI, pracovalo mnohem intenzivněji. Mezi možná nápomocná opatření patří tato:

- **Regulace interního využití:** Zavést komplexní monitorování toho, jak se systémy AI využívají v nejpokročilejších firmách, a omezit přístup k čistě nápomocným modelům „helpful-only“, které se bez omezení řídí jakýmkoli instrukcemi.
- **Otevřenost ohledně parametrů modelů:** Zveřejňovat podrobné informace o tom, jaké chování je v systémech AI zabudované, včetně pojmů a omezení, kterým jejich jednání podléhá. To umožní vnější kontrolu a odhalení možných chyb.
- **Širší sdílení kapacity:** Zajistit, že výkonné kapacity AI budou rozděleny mezi více zúčastněných stran a nebudou se koncentrovat v rukou několika jednotlivců nebo organizací. To vytvoří brzdy a protiváhy, díky kterým bude uzurpace moci obtížnější. Je však třeba vzít v úvahu, že příliš široké rozdělení výkonných kapacit AI také představuje rizika, takže je potřeba tyto protichůdné ohledy důkladně vyvážit.
- **Kontrola skryté loajality:** Vyvinout spolehlivé metody detekce, zda systémy AI nemají naprogramovanou skrytou agendu nebo zadní vrátka, která by jim umožňovala jednat způsobem odporujícím jejich udávanému cíli.
- **Pojistky u AI pro vojenské účely:** Trvat na tom, aby AI pro vojenské účely měla důkladné pojistky proti využití při pučích, včetně zákazu útočení na civilisty a požadavků na několik nezávislých schválení mimořádných akcí.

Pro mnohem více podrobností k tomuto tématu si poslechněte rozhovor s Tomem Davidsonem⁵.

Umělá inteligence představuje bezprecedentní riziko koncentrace moci. Na rozdíl od předchozích technologií může umožnit jednotlivci nebo malé skupině téměř absolutní kontrolu nad ekonomickými, vojenskými i správními systémy současně. Efektivní řešení vyžaduje mezinárodní spolupráci v následujících letech.

4 <https://www.forethought.org/research/ai-tools-for-existential-security>

5 <https://80000hours.org/podcast/episodes/tom-davidson-ai-enabled-human-power-grabs/>

Závěr

Vaše cesta k efektivnímu altruismu: Jak můžete pomoci co nejlépe

Dočetli jste se až na závěr Příručky efektivního altruismu. Doufáme, že vám poskytla užitečné mentální modely a nástroje pro systematické uvažování o tom, jak pomáhat.

Jak je v příručce popsáno, žijeme v éře nebyvalého potenciálu – poprvé v dějinách máme takovou *moc pomábat, a současně i takovou moc všechno zničit*. Efektivní altruismus (EA) je výzkumný projekt a zároveň světová komunita, která se snaží aplikovat poznatky získané rozumem a důkazy k tomu, abychom rozpoznali nejlepší formy konání dobra.

Tato příručka představila čtyři hlavní pilíře, na kterých úsilí efektivního altruismu stojí: stanovování priorit, nestranný altruismus, otevřené hledání pravdy a duch spolupráce. Vyzývá nás, abychom se nepoddali paralýze, která plyne buď z vidění jen toho, jak je svět hrozný, anebo jen toho, jak se zlepšil, ale abychom přijali, že svět *lze velmi zlepšit*.

Zároveň jsme se věnovali nejnaléhavějším problémům, které z těchto principů vyplývají, včetně *snižování existenčních rizik*, a to zejména rizik spojených s nekontrolovaným vývojem umělé inteligence (AI) a katastrofálními pandemiemi. Neopomenuli jsme ani globální zdravotní péči a chudobu, kde lze se stejným množstvím zdrojů pomoci stokrát, nebo i tisíckrát většímu počtu lidí než jinde.

Co je váš další krok? Efektivní altruismus neznamená jen studovat problémy, ale i aktivně aplikovat tyto poznatky do svého života. Je na nás všech, aby to dopadlo dobře. V sázce je budoucnost, která se může týkat nepsčtu generací. To činí ochranu budoucnosti tou největší světovou prioritou. Pusťte se do toho.

Vaše další kroky: Jak uplatnit efektivní altruismus v praxi

Efektivním altruismem se můžete řídit, ať už se chcete na konání dobra soustředit do jakékoli míry a v kterékoli oblasti života. Základní cesty, jak lidé jeho myšlenky uplatňují, jsou:

1. **Volba profesní dráhy:** Zvolte si profesi, která vám umožní podílet se na řešení naléhavých problémů, nebo najděte způsob, jak k tomu využít své stávající dovednosti. Klíčové je zaměřit se na mezní dopad – na to, jaký přidaný účinek bude

mít vaše činnost na danou oblast. Pro kariérní poradenství doporučujeme řídit se radami organizace **80,000 Hours**, která se specializuje na kariérní poradenství pro maximální dopad.

2. **Přispívání na charitu:** Darujte vybraným dobročinným organizacím, které dosahují nejvyšší nákladové efektivity. Pro výběr organizací s prokázanými výsledky doporučujeme využít výzkum organizací **GiveWell** nebo **Giving What We Can**.
3. **Budování komunity a zakládání organizací:** Podílejte se na budování komunit, které se věnují palčivým problémům (jako jsou rizika AI a pandemie). Můžete také zakládat nové organizace, které přispívají k řešení naléhavých globálních problémů, zejména těch opomíjených.

Doporučené zdroje

V češtině

Online:

- Další zdroje a odkazy na přednášky a kurzy naleznete na efektivni-altruismus.cz
- Novinky v AI můžete sledovat v newsletteru Pokrok v AI co píše Stanislav a Kristína Fortovi: stanislavfort.substack.com

V češtině vyšlo několik knih o těchto tématech; většina je dnes mírně zastaralá:

- **Nad propastí** — Toby Ord (2022, Argo). V originále: *The Precipice: Existential Risk and the Future of Humanity* (2020). Přehled existenčních rizik proč na nich záleží.
- **Dobré úmysly nestačí: Jak smysluplně pomábat díky efektivnímu altruismu** — William MacAskill (2021, Argo). V originále: *Doing Good Better* (2015). Hlavně o filantropii — jak se rozhodnout kam darovat.
- **Superintelligence** — Nick Bostrom (2017). V originále: *Superintelligence: Paths, Dangers, Strategies* (2014). Kanonické argumenty pro rizika ze ztráty kontroly nad AI. Napsáno před současným příchodem silných neuronových sítí. Filosoficky stále dobré.
- **Život 3.0** — Max Tegmark (2020, Argo & Dokořán). V originále: *Life 3.0* (2017). Speklativní mapování prostoru možných budoucností.
- **Jako člověk** — Stuart Russell (2021, Argo & Dokořán). V originále: *Human Compatible* (2019).
- **Praktická etika** — Peter Singer (2024, Karolinum). V originále: *Practical Ethics*, 3rd ed. (2011). Úvod do aplikované utilitaristické etiky.

V angličtině

- 80,000 Hours (80000hours.org) — Kariérní průvodce pro maximální dopad, zaměřeno na absolventy prestižních škol v UK a US. Obsahuje i seznam volných pozic s možným slibným dopadem: 80000hours.org/career-guide/
- Probably Good (jobs.probablygood.org) — Druhý seznam otevřených kariérních příležitostí. S širším záběrem než první.
- Třetí podobný seznam (a tady nejen pracovních ale i krátkodobých či dobrovolnických) příležitostí je na EA Opportunities board: na effectivealtruism.org/opportunities
- Dan Hendrycks — Introduction to AI Safety, Ethics, and Society (2024, Taylor & Francis, aisafetybook.com). Učebnice o bezpečnosti AI.
- Y. Bengio (chair) et al. — International AI Safety Report (2025). Mezinárodní

expertní soupis rizik frontier modelů, evaluačních mezer a doporučení pro stát i průmysl.

- What We Owe the Future by William MacAskill (2022). — Úvod do longtermismu: proč je důležité snížit riziko vyhynutí lidstva.
- Dwarkesh Podcast (dwarkesh.com). — Hluboké rozhovory s výzkumníky a zakladateli o AI a vědě.
- Don't Worry About the Vase (thezvi.substack.com) — Pravidelné analýzy současného dění v AI, píše Zvi Mowshowitz.
- Transformer Weekly: (transformernews.ai) — Týdenní přehledy dění v AI.
- effectivealtruism.org — Oficiální úvod do EA. FAQ.
- Effective Altruism Forum (forum.effectivealtruism.org) — Hlavní fórum k diskusi o efektivním altruismu.

Přednášky a akce

Zájemci o efektivní altruismus pravidelně pořádají akce (Praha, Brno, online) a víkendová setkání.

Přihlašte se k jejich sledování na:

- efektivni-altruismus.cz/kalendar-akci
- fb.com/efektivnialtruismus
- instagram.com/efektivni_altruismus
- Odebírejte newsletter na efektivni-altruismus.cz/newsletter

Zdroje kapitol

- *Co je to efektivní altruismus* bylo publikováno jako *What is Effective Altruism* na effectivealtruism.org/articles/introduction-to-effective-altruism
- *Jak hledat žlato* je zkrácená verze z *Prospecting for Gold*, který napsal Owen Cotton-Barratt na effectivealtruism.org/articles/prospecting-for-gold-owen-cotton-barratt
- *Mezní dopad* publikoval tým Probably Good jako *Marginal Impact: Making the Most of Additional Effort* na probablygood.org/core-concepts/marginal-impact/
- *Svět je brožný. Svět se velmi zlepšil. Svět lze velmi zlepšit* napsal Max Roser jako *The world is awful. The world is much better. The world can be much better* na ourworldindata.org/much-better-awful-can-be-better
- *Proč snižovat existenční rizika* je část z *The case for reducing existential risks* napsané Benjaminem Toddem na 80000hours.org/articles/existential-risks/
- *Prevence katastrofálních pandemií* je část z *Preventing catastrophic pandemics* napsané Cody Fenwickem a týmem 80,000 Hours na 80000hours.org/problem-profiles/preventing-catastrophic-pandemics/
- *Umělá inteligence mění náš svět – je na nás všech, aby to dopadlo dobře* napsal Max Roser jako *Artificial intelligence is transforming our world — it is on all of us to make sure that it goes well* na ourworldindata.org/ai-impact
- *Prevence katastrofy spojené s umělou inteligencí* napsali Benjamin Hilton a tým 80,000 Hours jako *Preventing an AI-related catastrophe* na 80000hours.org/problem-profiles/artificial-intelligence/
- *Uchvácní moci prostřednictvím umělé inteligence* napsal Cody Fenwick jako *AI-enabled power grabs* na 80000hours.org/problem-profiles/ai-enabled-power-grabs/

Online verzi příručky včetně všech odkazů naleznete na efektivni-altruismus.cz/prirucka.

Velice děkujeme původním autorům za svolení k překladu jejich textů.