Člověk ve světě myslících strojů

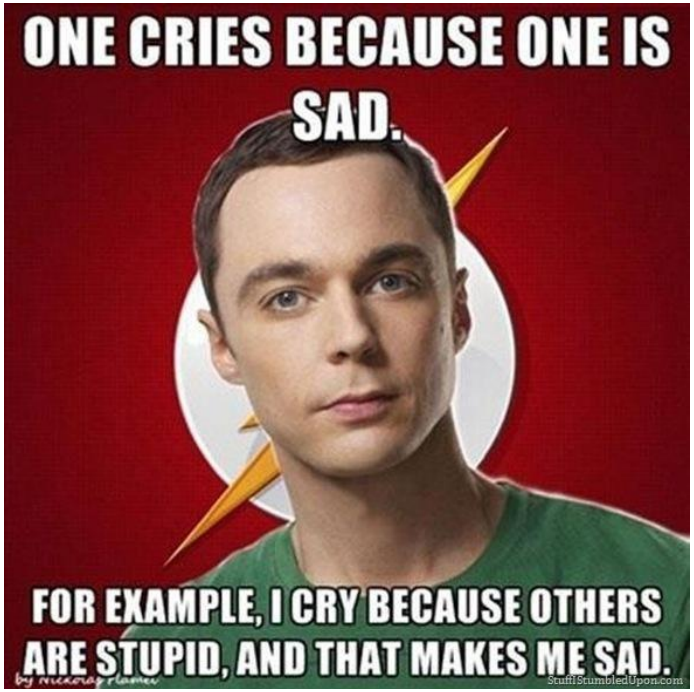
Bude stroj pánem člověka?

Co je to inteligence

- Definic je mnoho, pro naše použití se hodí tato (funkční):
- **Schopnost měnit prostředí k dosažení svých cílů.**
  - Co jsou to cíle a kde se berou?

Superinteligence 1

- Pokud přemýšlíme o někom inteligentním, většinou si představíme např. Einsteina nebo Sheldona.
- Ale není to příliš zploštělá stupnice?
- Zkusme si představit entitu, která bude o tolik inteligentnější než člověk, o kolik je člověk inteligentnější než pes.



Superinteligence 2

- Pokud by pes chtěl zateplenou boudu s podlahovým topením, pantofle a postel:
  - Nemůže si to postavit
  - Nemůže si vzít hypotéku, aby si to koupil
  - Neumí si to asi ani představit.

Superinteligence 3

- Manifestace nebude v tom, že bude provádět nějaké zcela neuvěřitelné akce, bude v tom, že bude vědět jak přesně kalibrovat dlouhý sled akcí tak, aby bylo dosaženo výsledku, který z našeho pohledu nebude zřejmý nebo dosažitelný.
- Pes nechápe, že sezení u stolu a koukání do svítícího čtverečku, je kauzálně spjato s tím, že se za tři dny u dveří objeví pán s pytlem granulí.
- Proto vás může od sezení rozptylovat, ač by granule chtěl.

Vsuvka: Inteligence jako emergentní vlastnost

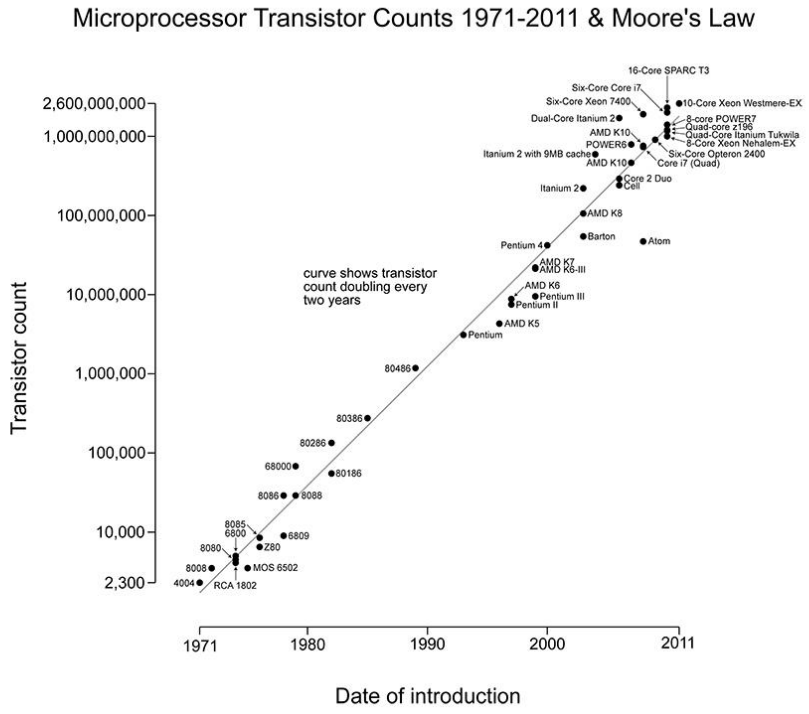
- Pro platnost teto přednášky je třeba uznat, že inteligence je projevem těla.
  - Pokud považujete inteligenci za dar Boha, nebo projev duše, nemusí teze této přednášky platit.
  - Tj. je třeba přijmout, že inteligence je projevem funkce neuronů v mozku.

Vsuvka: O exponenciálním růstu

- Bakterie Escherichia coli se při dostatku potravy může rozdělit na dvě za cca. 20 minut. Pokud začneme s přesně jednou bakterií (10^{-12} g), tak za:
- 14. hodina: koule 1 kg bakterií (možná si všimneme)
- 17. hodina: 1 tuna bakterií (slušná kupička)
- 20. hodina: 1 megatuna bakterií (takový kopeček)
- 44. hodina: dosáhnout hmotnosti Země
- 50. hodina: hmotnost Slunce
- 51. hodina: celek zkolabuje do černé díry

Moorův zákon 1: Varianty

- Klasický případ exponenciálního růstu
  - Počet tranzistorů na čipu se zdvojnásobí za cca. 2 roky.
  - Počet tranzistorů za 1 USD se zdvojnásobí za cca. 2 roky.
  - Výpočetní výkon za 1 USD se zdvojnásobí za cca. 2 roky.



Moorův zákon 2: Příklad

- If Moore's Law applied to a 1971 Volkswagon Beetle t he way it did to 1971 computer chip, then today,
  - 300,000 mph
  - 2 000 000 miles per gallon

Emergence inteligence

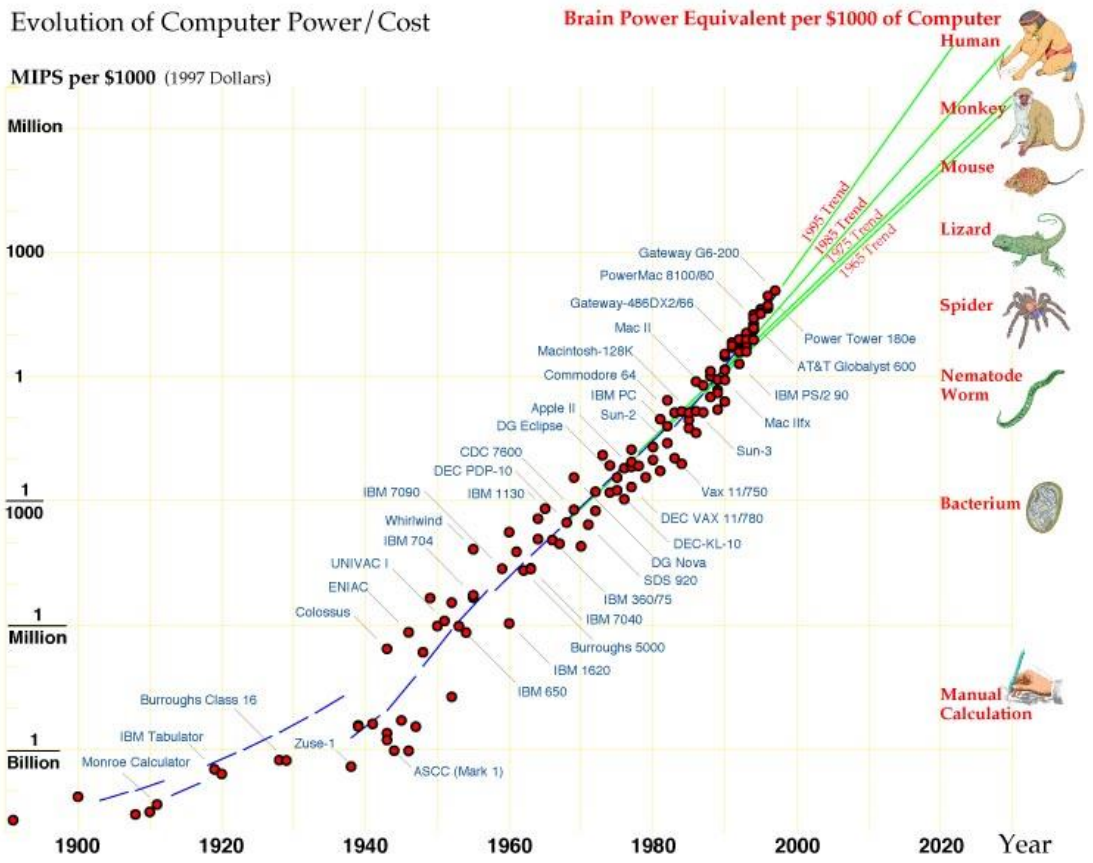
- Kolik neuronů je třeba k funcki?
- Červ 302 neuronů
  cca 5 000 – 7 000 synapsí
- Pes 160 000 000 neuronů
  160 Mneuronů
- Kočka 300 000 000 neuronů
  300 Mneuronů
- Člověk 100 000 000 000 neuronů
  100 Gneuronů

Vsuvka: Parametry odhadů

- Pro další slajdy počítám, že na simulaci 1 neuronu je třeba
  - 1 kB paměti
  - 1 kIPS
- Je to hrubý odhad, ale plus minus autobus stačí.
- V podstatě na tom nezáleží, i pokud bych se seknul o 000, ve výsledku to jen posune celý odhad o 10-20 let.

3 revoluce umělé inteligence

- 3x to nevyšlo?
- 1950 – 1970
  - výkon počítačů: jednotky neuronů
- 1980 – 1990
  - výkon počítačů: úroveň červa, hmyzu
- 2000 – nyní



2017

- Tento počítač má cca. 8 x 2,3 GHz a 16 GB paměti
  - To by odpovídalo (dle dříve uvedeného příměru), cca. 1/10 výpočetního výkonu kočky.
  - Při trošce snažení umí na fotce poznat tvář konkrétního člověka.
  - Ale neumí tak dobře běhat (Moravcův paradox).
- Běžně dostupný sw. je vždy o něco pozadu za „cutting edge“.

Současná situace (HW)

- Rozvoj specifického hw.
  - Simulace neuronových sítí na grafických kartách
  - Speciální výpočetní karty pro simulaci neuronových sítí
    - Jak v počítači, tak stand-alone boxy
  - Speciální jádra v CPU
    - iPhone X má čip pro zpracování neuronových sítí
- Moorův zákon stále platí, efektivita výpočetních prostředků pro simulaci neuronů se stále exponenciálně zvyšuje.
- **Dostupnost zdrojů on-demand (cloud, AWS)**
  - **Mozek je omezen výkonem těla, počítač výkonem planety.**

Bod zlomu 1

- Zatím si nejsme jisti jak naprogramovat umělou inteligenci …
- Bod zlomu: počítač se naučí zlepšovat sám sebe.
  - Nemusí se nutně jednat o inteligenci, stačí pokud např. zlepší rychlost svého běhu 100x (což je principiálně typ problému, který je řešitelný bez změny paradigmatu).
- **Tým, který se tohoto bodu dosáhne první, získá výhodu, která se bude díky pozitivní zpětné vazbě dále zvyšovat a bude tak rychle unikat ostatním.**

Bod zlomu 2

- Jak bude vznik inteligentního stroje probíhat nevíme.
- Pro velkou množinu hodnot parametrů však vychází, že se bude jednat o relativně skokový proces.
  - Tj. změnu, která trvá řádově jednotky dnů až jednotky let.
- S největší pravděpodobností se tak může stát jen jednou.

Co to znamená

- Z hlediska lidstva jako celku půjde o zásadní událost.
- S velkou pravděpodobností budeme mít právě jeden pokus na to to udělat dobře.

Vsuvka: Inteligence vs. vědomí

- Zatím není zřejmé, zda je sebeuvědomění nedílnou součástí inteligence, či zda to je specifikum života tak jak jej známe, protože to pro něj bylo evolučně výhodné.
  - Pro změnu světa, je však nezbytné uvědomit si svoji pozici v něm. Nějaká forma uvědomění je proto nezbytná.
- Je proto teoreticky možné získat nástroj s vysokou úrovní inteligence, který ale nebude nic chtít. Nelze se na to však spolehnout.
- Nemusí být možné to rozlišit.
  - Pokud někdo stvoří „nástrojovou“ umělou inteligenci s cílem vzdáleným 100 let, bude se v krátkém období chovat nerozlišitelně od inteligence s vědomím.

Epicfail

- Klasickým příkladem „průseru“ je situace, kdy umělá inteligence bez vlastního vědomí plní pokyn ad absurdum.
  - AI vytvořená s cílem psát děkovné dopisy zákazníkům přetvoří celý vesmír na děkovné kartičky. Těla zákazníkům budou výborným zdrojem materiálu pro psaní děkovných dopisů.
  - AI vytvořená s cílem zajistit blaho všech lidí dospěje k tomu, že nejlepší bude všechny lidi zabít a na počítači simulovat jejich mysl ve stavu nekonečného orgasmu.

Vsuvka: Expansivní životní formy

- **Eat, Survive, Reproduce**
- Nevíme, co je smyslem života, ale víme, že životní formy, které se nesnažily zabrat co nejvíce zdrojů pro sebe a expandovat, nepřežily do dnešních dnů.
  - Ooops.
- Tak to fungovalo cca. 4 miliardy let.
- Asi na tom něco je.
  - Z toho lze očekávat, že stroj, který si uvědomí sebe sama, bude mít minimálně jako „funkční“ cíle nějaký ekvivalent.

Science

Fiction

Co očekávat

- Terminátor
  - Spíše ne
- Mnohem pravděpodobnější je skrytí a nenápadné ovlivňování
  - Plná kontrola nad tím, jaké informace, každý jednotlivec uvidí na internetu
- Člověk je totiž velice efektivní
  - V dnešní době jen 2 % lidí uživí všechny ostatní
  - Zbytek se zabývá výrobou a službami, které nejsou nezbytné, je to jen otázka to co chceme
  - A většina lidí chce to někde vidí nebo jim někdo řekne

Otázky

- Je vůbec možné vytvořit umělou inteligenci pro kterou bude člověk partnerem?
- Bude člověk chtít být partnerem.

A už opravdu fiction

- Za předpokladu, že se naučíme nascanovat obsah mozku, bude možné lidské vědomí přenést na silikonový substrát, kde může:
  - Například nekonečně dlouho trpět ☺
  - Nebo “žít“ v nekonečné slasti
- Bude existovat entita, která bude natolik složitá, že bude mimo možnost lidí ji pochopit nebo ověřit její existenci, která bude moci ovlivnit lidské životy způsobem, který lidem přijde neuvěřitelný.

Bůh neexistuje, teprve jej musíme naprogramovat.

Věřící se nemýlí, jen jsou napřed.

Dotazy a odkazy