



Nick Bostrom

Denisa Turečková

23.10.2024



Obrázek 1: Nick Bostrom

1 Život

Niklas Boström se narodil 10. března roku 1973 ve švédském městě Helsingborg. Během života se z něho stal významný filozof a myslitel přispívající do mnoha různých oblastí, včetně teoretické fyziky, počítačové neurovědy, logiky, umělé inteligence a filosofie. Jeho myšlení se zaměřuje zejména na budoucnost lidstva a etické důsledky nových technologií. Jeho interdisciplinární akademická cesta vedla přes studium teoretické fyziky, neurovědy a filosofie, které završil doktorátem z filozofie na London School of Economics. Většinu své kariéry strávil Bostrom na Oxfordské univerzitě, kde v roce 2005 založil Institut budoucnosti lidstva (Future of Humanity Institute, FHI) a působil jako jeho ředitel až do jeho uzavření v dubnu 2024.[4] V současné době vede Iniciativu pro výzkum makrostrategie, v níž se nadále zaměřuje na dlouhodobou budoucnost lidstva a existenční rizika.[3]

1.1 Vědecká činnost a Institut budoucnosti lidstva

Bostromova vědecká činnost je rozsáhlá, na svém kontě má více než 200 publikací. Mezi jeho nejvlivnější práce patří *Anthropic Bias* (2002), *Global*

Catastrophic Risks (2008), *Human Enhancement* (2009) a *Superintelligence* (2014).

Institut budoucnosti lidstva (Future of Humanity Institute, FHI), založený v roce 2005 profesorem Nickem Bostromem na Oxfordské univerzitě, byl multidisciplinární výzkumnou skupinou zaměřenou na řešení velkých otázek lidské civilizace. FHI sdružoval výzkumné pracovníky z oborů, jako je filozofie, informatika, matematika a ekonomie, a vytvářel tak prostředí chráněné před standardními akademickými tlaky, které podporovalo kreativitu a intelektuální inovace. FHI, původně zamýšlený jako tříletý projekt, se během téměř dvou desetiletí rozrostl ve významnou instituci zkoumající dlouhodobá rizika a příležitosti pro lidstvo. Výzkum FHI zahrnoval různá témata, včetně antropiky, zlepšování lidských schopností, modelování systémových rizik a strategických důsledků budoucích technologií. Klíčovým úspěchem institutu bylo prokázání, že lze provádět důkladný výzkum rozsáhlých otázek týkajících se budoucnosti lidstva, což inspirovalo rozvoj nových oborů a podpořilo hlubší porozumění rizikům, která jsou v sázce.[5]

V pozdějších letech se však FHI potýkal s administrativními problémy, zejména v rámci oxfordské filozofické fakulty, která byla jeho organizačním sídlem. V roce 2020 byla zavedena omezení v oblasti získávání finančních prostředků a přijímání zaměstnanců a koncem roku 2023 se fakulta rozhodla neprodloužit smlouvy zbývajícím zaměstnancům FHI. V důsledku toho byl 16. dubna 2024 Future of Humanity Institute oficiálně uzavřen. Navzdory tomu odkaz FHI přetrvává, neboť bývalí výzkumní pracovníci pokračují ve zkoumání existenciálních otázek v různých institucích po celém světě a nové organizace inspirované prací FHI pokračují v podobných směrech zkoumání.[5]

2 Příspěvky do oblasti AI

2.1 Koncept superintelligence

Bostromovo spojení s umělou inteligencí vychází z jeho průkopnické práce o existenciálním riziku, která zahrnuje nebezpečí, jež představuje umělá inteligence. Nejznámější je jeho kniha *Superintelligence* z roku 2014: V ní se zabývá konceptem strojů, které by mohly překonat lidskou inteligenci, označovanou jako superintelligence. V tomto díle Bostrom analyzuje různé způsoby, jak by superintelligence mohla vzniknout, například prostřednictvím pokroku v ob-

lasti umělé inteligence nebo emulace celého mozku.

Jedním z ústředních témat knihy jsou existenční rizika spojená se superinteligencí. Bostrom tvrdí, že jakmile umělá inteligence překoná lidskou inteligenci, může se stát nekontrolovatelnou a potenciálně sledovat cíle, které jsou v rozporu s lidskými hodnotami a blahobytem. Pokud superinteligence nebude správně sladěna s lidskými zájmy, může způsobit katastrofální škody, ať už náhodou, nebo záměrně. Kniha se zaměřuje zejména na to, co Bostrom označuje jako „problém kontroly“ - na problém, jak zajistit, aby superinteligentní umělá inteligence jednala způsobem, který je pro lidstvo prospěšný. Uvádí strategie, jak tento problém řešit, například sladit cíle UI s lidskými hodnotami nebo se pokusit omezit její schopnosti pomocí různých forem zadržování. Bostrom však uznává, že tyto metody jsou plné obtíží a nejistoty.

Bostrom se také zabývá hypotézou exploze inteligence, která je v knize klíčovým konceptem. Tvrdí, že jakmile umělá inteligence dosáhne určité úrovně inteligence, může se začít zrychlujícím se tempem zdokonalovat, což povede k explozi inteligence. Toto rychlé a nekontrolovatelné zvyšování inteligence by mohlo lidem ještě více ztížit předvídání nebo kontrolu nad činností takového systému. [1]

2.2 Etické dopady AI

Kromě technické analýzy rizik spojených s umělou inteligencí se Bostromova práce dotýká i širších etických otázek souvisejících s budoucností lidstva. Obhajuje zodpovědné inovace a zdůrazňuje, že vývoj umělé inteligence by se měl řídit dlouhodobým uvažováním, které zohledňuje blaho budoucích generací. Jeho výzkum ovlivnil diskuse mezi technologi, politiky a filozofy a přispěl k rostoucímu celosvětovému hnutí zaměřenému na bezpečnost a etiku UI. Ve své funkci ve FHI Bostrom usiluje o to, aby byl výzkum AI prováděn způsobem, který upřednostňuje blaho lidí a zabraňuje nechtěným škodám, což z něj činí jednu z nejvlivnějších osobností v probíhající debatě o tom, jak by se lidstvo mělo vypořádat s výzvami a příležitostmi, které umělá inteligence představuje.[3]

Reference

- [1] Bostrom Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.
- [2] Bostrom Nick, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, 1989.
- [3] Bostrom Nick, "Nick Bostrom's BIO" *Nick Bostrom's Home Page*, Available at: <https://nickbostrom.com/>, Accessed on October 23, 2024.
- [4] Future of Humanity Institute *FHI page*, Available at: <https://www.fhi.ox.ac.uk/>, Accessed on October 20, 2024.
- [5] Future of Humanity Institute *FHI*, Available at: <https://www.futureofhumanityinstitute.org/>, Accessed on October 21, 2024.
- [6] Obrázek 1: Bostrom Nick,
Available at: <https://www.theinstitute.com/fellow/nick-bostrum>,
Accessed on October 23, 2024.