novinky.cz/kultura/salon; facebook.com/salonpravo

Všichni věřili,
že toto ještě není
konec světa,
protože takhle
konec světa
přece nevypadá.

Emil Hakl
v knize *Konec světa* (2001)

Salon
PŘÍLOHA PRÁVA
1089
13. září 2018

Ivan M. Havel o strachu ze superinteligence:

Jen další „konec světa“

Štěpán Kučera

Ivan M. Havel (1938), vědec a zakladatel Centra pro teoretická studia, se už od sedmdesátých let zabývá umělou inteligencí a filosofickými otázkami s ní souvisejícími. Salon se ho proto zeptal na téma, o němž dnes někteří píší s obavami a někteří s nadějí – vznik tzv. superinteligence.

Foto Jan Handrejch

S Jak vzniká lidské myšlení?

Rád bych nejprve rozlišil dvě hlediska, jak chápat lidské myšlení. Zaprvé o něm můžeme mluvit jakoby z vnějšího, nezaujatého stanoviska, podobně jako mluvíme o jakémkoliv přírodním ději – třeba o tom, že za oknem padají kroupy. A zadruhé lze o myšlení mluvit jako o něčem, co vědomě prožívá ten, kdo zrovna myslí – například já, když vám říkám tuto větu. Kroupy si prostě letí věrny přírodním zákonům a nikoho nenapadne prohlašovat, že při tom myslí. Já jsem na tom jinak: pečlivě volím slova a snažím se, aby věta dávala přesně ten smysl, jaký jsem zamýšlel. Prostě jsem u toho a vím o své snaze a své svobodě volit slova. Jinak řečeno vím, že jsem, že se o něco pokouším a že při tom myslím. Abychom si dále rozuměli, budu první hledisko označovat – nepřesně – jako *vnější* a druhé – opět nepřesně – jako *vnitřní*.

> Je jistě dobré, abychom se na leccos předem připravili, ale jak se můžeme připravit na *všechno*?

Myslím, že podstatné je právě ono druhé, *vnitřní* hledisko, které se nedá jen tak odmávnout. Nedovedu si představit lidské myšlení bez přihlédnutí k tomu, *kdo* myslí, přesněji řečeno, kdo si *uvědomuje*, že myslí, a že ten „kdo“ jsem třeba já.

Ptáte se, jak lidské myšlení vzniká. Beru to jako otázku po jakési „hnací síle“ v průběhu myšlení, a pak mě předchozí rozlišení dvou hledisek navádí na dvě odpovědi. Z *vnějšího* hlediska bychom se asi dostali k otázce po biologické povaze mentálních procesů. Dnešní poznatky neurověd – od elektrochemických dějů na synapsích až po funkční konektivitu rozsáhlých oblastí mozku – jsou obdivuhodné, nicméně odhalují nanejvýš jen určité korelace mezi děním v mysli a děním v mozku. Neříkají nic o kauzálním vztahu mezi nimi. Takže vědecky přijatelná odpověď na otázku po vzniku myšlení zatím chybí.

Z *vnitřního* hlediska je však situace ještě tristnější, ale právem. Pakliže je pro člověka charakteristické, že si plně uvědomuje své myšlení, a tím i sebe sama v roli myslícího, že ví o své osobní identitě, vnitřní svobodě a o svých tužbách i obavách, jde o něco, co se vymyká jakémukoliv objektivistickému popisu, neřkuli experimentu. Svůj osobní přístup k sobě samému chápu jako ontologicky jinou sféru zkušenosti, než je zkušenost s vnějším světem.

S A kde začíná člověk? Co nás odlišuje od zvířat?

Abych dodržel svou preferenci *vnitřního* hlediska, rád bych dospěl k charakterizaci člověka nejen jako myslící bytosti, ale jako *vědomě* myslící bytosti, takové, která o sobě především nějak ví. Chápeme-li svět dějinně – v tom smyslu, že věříme, že svět se nějak vyvíjel nebo alespoň měnil v čase –, stojíme před nezvratnou skutečností, že byly doby, kdy člověk na naší planetě chyběl. Muselo tedy existovat období, kdy někteří, asi i jinak vyspělí živočichové takříkajíc vnitřně „prozřeli“. Nikdy se nedozvíme, jak dlouho to trvalo a jaké to pro ně bylo.

Pamatujete si na film *2001: Vesmírná odysea*? Je tam na začátku dlouhá pasáž, v němž se tlupa lidoopů mění v tlupu opolidí. Explicitně je ukázáno, že v tom hraje roli objev kosti coby nástroje k lovu i k obraně. Já bych však zdůraznil něco jiného: v tom filmu vznikala sehraná tlupa a její souhra byla založena na vzájemné komunikaci pomocí skřeků a gestikulace. Ke vzniku člověka tam došlo v tlupě, nikoliv v jedinci.

Zatím jsem tvrdil, že hlavní charakteristikou člověka je vědomí sebe a své vlastní identity. Kde se však toto vědomí a sebeuvědomění vzaly? Není to tak, že k tomu musí existovat i druzí, kteří se ke mně jako k jedinci obracejí, navazují se mnou kontakt, berou mě do tlupy a – koneckonců, ale hlavně – se se mnou domlouvají nějakou řečí?

Užívání řeči je mimo jiné dokladem, že s námi, lidmi, je to nějak jinak než u zvířat. Mezi lidmi si vzájemně rozumíme proto, že shodně prožíváme svět, a nic bychom shodně neprožívali, kdybychom spolu nemluvili.

S V čem se podle vás myšlení umělé inteligence může podobat člověku? Ptám se samozřejmě hlavně na vědomí sebe sama.

Snad všechny dosavadní dovednosti umělé inteligence obsahují něco podobného lidskému myšlení. Počítače umějí hledat cestu v bludišti, rozpoznávat lidské tváře, obstojně překládat mezi jazyky, řídit auta… Mnohé z toho se dá kvantifikovat co do úspěšnosti a pak většinou počítače vyhrávají nad člověkem – hlavně díky své operační rychlosti a kapacitě paměti. A právě tyto úspěchy vedly ke zrodu dvou velkých idejí s dalekosáhlými důsledky.

První je idea obecné umělé inteligence. Všechny konkrétní dovednosti, nejen ty, které jsem uvedl, ale mnohé myslitelné další, lze spolu prostě sečíst. Vytvořit jakýsi všeumělý počítač. A právě od toho se odvíjí idea druhá, idea superinteligence, jakési „nadlidské“ inteligence. Představte si obecnou umělou inteligenci, která navíc dalece překonává schopnosti lidské ve skoro všech dovednostech. Tak to alespoň definují autoři v čele s Nickem Bostromem.

Nad představou superinteligence se výrazně ohlašuje problém, na nějž jste se zeptal, totiž vědomí sebe sama. Přesněji řečeno, to, co se ohlašuje, je jeho ignorování. Příkladem je zmíněný Bostrom, který už někde na straně 47 své objemné knihy *Superinteligence* předesílá, že otázkou, zda by superinteligence mohla mít subjektivní vědomou zkušenost, se zabývat nebude.

Jak jsem už uvedl, lidské myšlení se nedá oddělit od toho, že vždy je tu někdo, *kdo* myslí a *kdo* si své myšlení, a tedy i *sebe sama*, uvědomuje. Má-li být umělá inteligence aspoň v něčem odvozena od inteligence lidské, nebo ji dokonce předčít, pak je nutné si tuto otázku položit. Zdálo by se, že lze uhnout pohledem a vsadit na hledisko, které jsem nazval *vnější*, totiž že za vším jsou jen fyzické, kauzální procesy. Pak bych měl ovšem velký problém s připisováním některých typicky niterných vlastností superinteligenci, jak to opakovaně činí Bostrom. Příklady: „vylepšovat sebe sama“, „chovat se přátelsky“, „mít své motivace“, „chtít získat moc“.

S Bostrom přirovnává filosofy k psům, kteří se snaží chodit po zadních, a vyslovuje hypotézu, že člověk možná nemá na to, zodpovědět základní otázky lidstva. Podle něj k tomu superinteligence může být způsobilejší.

Naštěstí jsem tuhle pasáž přeskočil. Považuji to za spekulace, které ani moc nemá smysl kritizovat. Pokud ten jeho geniální superpočítač o sobě nebude vědět, nebude ani vědět, že o sobě víme my. A vůbec, jaké základní otázky? Rád bych viděl, jak by je superinteligence formulovala. Už to by si vyžádalo hlubokou orientaci v dějinách filosofie. A vůbec nevím, proč bychom měli čekat na nějakou superinteligenci, aby za nás všechno vyřešila…

S Bostromovi šlo v knize asi hlavně o to, upozornit na bezpečnostní rizika, která s sebou může vývoj umělé inteligence nést.

V současné době zažíváme inflaci zájmu o následky překotného vývoje umělé inteligence. Zdá se mi, že tato inflace není jen důsledkem té překotnosti, ale že je i projevem obecné přitažlivosti dystopických vizí konce světa. Co kdysi znamenala srážka s kometou, může dnes být superinteligence. Co na tom lidi tolik vzrušuje? Že by vidina, jak v poslední chvilce před zánikem světa se lidi semknou a začnou se navzájem mít rádi?

Ať tak či onak, způsob, jak se o budoucnosti píše, na mě dělá dojem, že autoři jsou zajedno se čtenáři v tom, že superinteligence bude nebezpečná a že nás, jako inteligenčně zakrnělé tvory, dříve či později vypudí ze světa. Nikdo se neptá, proč by to tu superinteligenci mělo napadnout. Je jistě dobré, abychom se na leccos předem připravili, ale jak se můžeme připravit na *všechno*?

S Bostrom netvrdí, že bude superinteligence zlá. Nabízí třeba myšlenkový experiment s výrobou kancelářských svorek – superinteligence dostane zadání je vyrábět a postupně přemění celý svět v montážní linku na výrobu svorek, čímž mimoděk vyhladí lidstvo.

Tak to by opravdu bylo pitomé zadání… V tomhle příkladu jde ovšem Bostromovi o jinou věc: o to mít na paměti, že zkáza by mohla vzniknout jako vedlejší, instrumentální produkt na první pohled neškodného záměru. Bostrom navrhuje takovým úletům zamezit vytvořením jakéhosi systému hodnot pro superinteligenci, který by respektoval i lidské hodnoty. Nebo by si takový systém vytvořila sama, inteligentní na to bude dost. Ale o jaké „hodnoty“ by šlo, když ten počítač nemá ponětí, co je dobro a co zlo?

S Vývoj umělé inteligence pokračuje, nezávisle na tom, jestli si to my dokážeme představit. Probíhá v různých soukromých firmách, o lecčems ani nemusíme vědět. Když nesdílíte katastrofické scénáře, tak jak to podle vás bude dál?

Nejde o to, jestli je sdílím, ale zda je vyhledávám. Různé katastrofy máme za rohem, i letošní parné léto nám cosi připomnělo. To, jak Bostrom mluví o superinteligenci, se mi zdá dost vágní, protože libovolné. Jak může superinteligence něco opravdu *chtít* – třeba nás vyhubit? Vždyť ani nevíme, co bude o sobě vědět! Co když si bude myslet, že je člověkem, a vyhubí se sama?

Vy se ovšem ptáte, co bude dál. Myslím, že nastane ohromný pokrok v konkrétních, řekněme užitečných směrech. O to se ty firmy bezpochyby postarají, avšak nějak se mi nechce věřit, že jim půjde o rozvoj zcela obecné umělé inteligence…

Ostatně po tom letošním vedru by mi stačilo, kdyby se nějaká hluboká neuronová síť naučila líp předpovídat počasí. To je jedna z mnoha úloh, na které by umělá inteligence mohla využít svých skvělých schopností a také toho, že neví, co je to únava či lenost.

A na závěr: i já mám své znepokojení a poradit si s ním neumím: Pokud si zvykneme na superinteligenci a na její umělé myšlení, co to udělá s námi?

krácená verze, celý rozhovor už příští středu
na **novinky.cz/kultura/salon** a **facebook.com/salonpravo**