Empatie robota

Empatie, čili schopnost vcítit se do emocí, nálady nebo citových stavů druhého, je základním znakem vztahů mezi normálními lidmi. Může být povrchní a jaksi automatická vůči neznámým lidem, nebo hluboká až na hranice identifikace s těmi, které milujeme. Neurofyziologie nás poučuje, že základy primitivní empatie jsou dány už v anatomii našeho mozku: máme prý každý tzv. *zrcadlové neurony*, díky nimž se v našem vědomí jako vnitřní pocit odráží manifestní bolest, radost nebo úzkost případně vztek druhého. Citlivost lidí se v tom směru liší, tak jak už různá bývá vůbec jejich vnímavost – u některých psychopatů může být až téměř zcela potlačena, u jiných, přecitlivělých lidí může vést až k citovým otřesům.

Empatie, jak se běžně projevuje, jistě nezávisí jen na specializovaných neuronech, ale je součástí složitého citového života, který závisí na výchově, předchozí zkušenosti a také do značné míry na vývoji kultury. Je pro nás dneska sotva představitelné, že před několika stoletími lidé běžně přihlíželi veřejným popravám, jejichž brutalitu člověk ani nechce popisovat. Obliba literatury a filmů, v nichž se líčí nebo přímo zobrazují krajní krutosti, naznačuje, že fantazie mnohých z nás není té necitlivosti středověkých lidí až tak vzdálená, jakkoliv se zaštiťuje oním modem *jako by*, vzdáleným skutečnosti, v které žijeme.

V běžném životě ale skoro každý má někoho blízkého, s kým cítí, tedy pociťuje utrpení nebo alespoň duševní tíseň při vědomí jeho bolesti, sdílí s ním jeho radost stejně jako oprávněný smutek, dovede se vžít do jeho úzkostí a podobně. A skoro každý má – nebo alespoň touží mít – někoho, kdo podobně cítí s ním. Ano, přehnaná empatie může někdy obtěžovat, zvlášť když své vlastní city potlačujeme nebo se za ně stydíme a rádi bychom je před ostatními skryli, ale to jsou spíše výjimečné případy: mnohem častěji je nám příjemné, když můžeme své emoce s někým sdílet. Často si o to říkáme, když si například stěžujeme na své bolesti či potíže, nebo naopak sdělujeme radost nad nějakým úspěchem či kladné vytržení nad zážitkem krásy a tak podobně.

Je smutným faktem, že díky vývoji životního stylu posledních několika generací a také prodlužování života přibývá osamělých lidí, z nichž někteří se nemohou o sebe plně postarat. Domovů, které by jim zajistily péči a překonaly jejich osamělost aspoň povrchně, už dávno není dost a je v nich – nebo mimo ně – stále citelnější nedostatek ošetřovatelů. Myšlenka, že by je v budoucnosti mohli nahradit roboti, není zdaleka nová – je tu s námi alespoň půl století. Jejímu uskutečnění napřed bránily nejen city (zdálo se téměř hrůzné, že by o člověka měl pečovat stroj), ale i nedostatky potřebné techniky. Jak to už ale bývá, pod tlakem potřeby a proměnami mentality zejména v Asii se odpor k této možnosti změnil v překonatelný předsudek a zároveň technika jak v oblasti robotiky tak zejména umělé inteligence udělala nečekaný skok. V zásadě máme dneska mechanismy, které se umí šetrně a přesně starat o základní životní funkce bezmocného pacienta, aniž by jejich výkonu bránil odpor, netrpělivost a vyhasnutí zájmu, které se často vyskytují u živých ošetřovatelů. S jistými výhradami se dá říct, že alespoň v laboratořích už existují systémy, schopné analyzovat řeč nejen na hlasové, ale i na sémantické úrovni a na základě pokročilé statistiky větných spojení věrohodně konverzovat s člověkem pomocí hlasového syntetizátoru. Jistě lze v obou směrech ještě

mnohé vylepšovat, ale zdá se, že je už nedaleko chvíle, kdy bude možné postavit robota, který nejen bude schopen pomáhat, ale dělat člověku i společnost a tím překonávat jeho osamělost. Až na jedno, co donedávna mezi odborníky nikoho nenapadlo: inteligentní robot, jak jsme ho tu načrtli, není schopen žádné empatie a jeví se proto jako *bezcitný*.

Nabízí se ovšem námitka, že robot žádné city mít nemůže, tedy se jako bezcitný nejen *jeví*, ale také nutně *je*. To je zřejmě pravda, alespoň prozatím – otázka, zda je možné na neživé bázi vytvořit cosi jako vědomí s jeho různými projevy, zůstává otevřená; odpověď na ni, ještě v nedávné době samozřejmě záporná, se trochu znejišťuje pokroky umělé inteligence na poli vnímání a myšlení. Rozpoznání složitých vzorců, například lidské tváře, na něž jsme my lidé byli vrozeně jedinými experty, je dnes pro umělá zařízení samozřejmostí; porovnání například otisků prstů s tisíci vzorů, ale také lidských genomů, strukturních prvků hmoty, tvarů milionů galaxií, to jsou jenom některé příklady schopností, v nichž už inteligentní stroje vykazují neselhávající přesnost a zejména rychlost, s níž se lidské vnímání už nemůže měřit. Zvykli jsme si dávno na to, že naše často i stolní počítače provádějí výpočty, které svou složitostí a zejména počtem operací v reálném čase daleko přesahují lidské možnosti a už se bez nich proto nemůžeme obejít. Říkáme si přitom ovšem, že ty stroje *nevědí, co dělají*, prostě jen velmi rychle vykonávají operace, předepsané programem, vytvořeným člověkem. Tak tomu jistě často je. Jistota v tomto směru byla ale už před lety narušena, když si počítače – jistě na prvotní lidský popud – začaly samy navrhovat své složité obvody, když se jejich programy začaly učit ze svých chyb a samy se opravovat, když dokázaly zdůvodňovat (jistě v příslušném kódu) nová řešení, která navrhly. Dnes už jsou stále lépe vyvíjeny programy, které dokážou v symbolickém jazyce provést matematický důkaz nějakého tvrzení – a nikdo neví, kde jsou meze takového vývoje.

Inteligentní robot, o němž byla shora řeč, bude se mnou věrohodně konverzovat, nikoliv ovšem na základě skutečného *rozumění*, jakkoliv působivou iluzi ve mně svými replikami vyvolá. Nicméně uvažme to trochu, než postoupíme dál. Co vlastně myslím tím, když řeknu, že mi můj protějšek *rozumí*? Snad mi tu bude dovoleno uvést příklad pro to, že ta otázka zdaleka není banální. Pozorujeme občas s mou ženou na procházkách sympatického mladíka a jeho psa, jejichž vztah je těžké nazvat jinak, než jako *srozumění*: pán na psa nikdy nekřičí, téměř mu ani nedává nějaké příkazy, nechává mu značnou volnost v jeho hrách a přitom je na něm znát klidnou jistotu, že pes na něho počká tam, kde má, a podobně. Nedávno jsme však byli svědky zvláštní příhody. Pes zachytil v letu nějakou hračku (plastický disk), s níž si hráli cizí lidé. Nějakou chvíli si s ní hrál, pak ji přinesl v tlamě svému pánovi. Ten mu klidně řekl: „Tak to už stačí. Teďka jdi a vrať těm lidem tu hračku, potom půjdeme domů.“ Pes bez váhání udělal, co pán žádal – odnesl hračku majitelům a pak se vrátil k pánovi. Bezděky jsme oba vykřikli: „Ten pes vám rozumí!“ Mladík spokojeně kývnul a se psem odešel.

Čím byl ten příklad zvláštní? Většina psů, které v parku vídáme, patrně správně reaguje na krátké povely jako „Sedni; ke mně; zůstaň“ a tak podobně. Váhám říkat, že těm povelům *rozumějí*, jelikož vídám někdy zdlouhavou proceduru, kterou jsou k správné reakci cvičeni. Když pes potom takový povel poslechne, lze to brát téměř jako podmíněný reflex ve smyslu Pavlova. Dojem, že doopravdy *rozumí*, jako v tom uvedeném případě, vzniká zřejmě tím, že situace asi nebyla tak běžná, pokyn byl složitější a té reakci nepředcházel dlouhý dril.

Přesto je těžké, být si doopravdy jist. Když mám – většinou oprávněný – dojem, že mé promluvě nějaký člověk *rozumí*, předpokládám se samozřejmostí, že jeho adekvátní reakce je vyvolána shodným nebo aspoň sourodým mentálním obrazem s mým vlastním, jejž jsem svými slovy vyjádřil – jinak řečeno, že sdílíme zhruba tentýž *význam*. O tom, co ten *význam* je, bylo už mnoho napsáno, což ale neznamená, že tím bylo odhaleno celé jeho tajemství. Mohu samozřejmě – v souhlasu s pozorováním svým i mnoha jiných – říct, že je to trojstranný vztah, v němž k něčemu (jevu, jsoucnu, pojmu) poukazuji někomu (třeba jenom potenciálnímu) jinému, případně reflexivně jen sám sobě. Ve skutečnosti je to mnohem složitější: význam skoro vždy vystupuje v nějakém kontextu, který jej propojuje s jinými významy, z nichž některé (předchozí a spolumíněné) jej vyvolaly a jiným se jako možným otevírá; kromě toho význam tak říkajíc obklopují různé konotace a možné asociace, některé zjevné, jiné skryté. To všechno význam zabarvuje a způsobuje jeho *život*, tedy jeho jemné proměny a hlavně schopnost tvořit půdu pro nové významy. Nejenom (jak někde poznamenal Bertrand Russell), že si vzájemně přesně nerozumíme (a proto prý můžeme komunikovat): v určitém smyslu nerozumíme tak docela ani sami sobě, čímž chci jen říct, že své významy stále trochu hledáme a dotváříme. Nicméně tento trochu unikavý význam přece jenom *mám*, je to *můj* význam, potvrzený mým subjektem – jinak řečeno, *uvědomuji si jej* (byť třeba ne úplně v jeho možné šíři nebo hloubce).

Význam kromě toho není jenom spojka mezi mnou, protějškem a nějakým jevem (a jak jsme řekli, v neurčité sítí s řadou dalších souvisejících významů): je to něco víc, co dané jsoucno nebo jeho pojem jaksi vyzdvihuje, dodává mu váhy (v té souvislosti je namístě připomenout gradaci účel-význam-smysl). Ne náhodou existuje ve všech evropských jazycích od *významu* posun k slovu (nebo pojmu) *významný*, označujícímu jistou existenciální naléhavost nebo váhu. I význam sám nese určitou závažnost, nějak se mne a případně jiných *týká*. To mluvím o živých lidských bytostech. Jak je to ale s robotem?

Když promluvím na některého z relativně dokonalých robotů, jaké se staví v předních laboratořích umělé inteligence, dojde (podle dostupných zpráv) zhruba k těmto dějům: nejprve jeho analyzátor řeči ve zvucích mé promluvy rozpozná jednotlivé fonémy, z nich složená slova a jejich větná spojení; to vše potom jiný modul softwaru porovná s jeho pamětí, určí, zda jde o příkaz (který robot provede) nebo konverzaci; ve druhém případě statisticky vyhodnotí možná slova resp. věty, které se mohou vyskytovat ve spojení s těmi, jež jsem pronesl, a tu nejpravděpodobnější možnost předá modulu hlasového syntetizátoru, který zařídí, že z amplionu zazní odpověď. Jelikož všechny tyto moduly jsou vytvářeny tak, aby se učily ze svých nezdarů (zejména eliminací neúspěšných spojení), robot po mnoha desítkách hodin *tréninku* je schopen se mnou inteligentně konverzovat o všem možném, aniž by něco porušilo dojem, že se mnou opravdu *mluví*. Znamená to, že mi také *rozumí*?

To není tak banální otázka, jak by se mohlo zdát. Pro Derridu a jeho následovníky by naše konverzace byla prostě interakcí *textů* a otázka by vlastně postrádala smysl. Pro extrémního empiristu by také byla prázdná, jelikož zjevně neexistuje způsob, jak se o jejím předmětu přesvědčit. Krajní logický pozitivista by konstatoval, že pokud robot na všechny mé věty reagoval logicky adekvátním způsobem, není vlastně, nač se ptát. Žádná z těchto reakcí pro nás patrně nemůže být uspokojivá. Než se pokusím naznačit svou vlastní odpověď, dovolím si dát krátký příklad ze svého dětství.

Patřil jsem v jistém věku k dětem, které někteří lidé označují jako *přemoudřelé*, což není příliš vlídná charakteristika. Vládl jsem na svá léta velkou slovní zásobou, bystře jsem uměl pochytit některé úvahy či fráze dospělých a taky je opakovat v správném kontextu, takže jsem se myslím často nezesměšnil, i když mé věty někdy vyvolaly udivené úsměvy. To, že jsem často mluvil jako dospělý, bylo patrně hlavně proto, že jsem toužil potěšit své rodiče a vysloužit si jejich pochvalu. Podstatné ale bylo, že jsem někdy jenom sotva tušil, co ty výroky znamenají (neměl jsem ještě k tomu vzdělání a hlavně odpovídající zkušenost). Až potud by se dalo říct s určitou nadsázkou, že jsem se v tom směru choval jako inteligentní robot. Brzy jsem ale nějak začal pociťovat falešnou dospělost svých reakcí a to mě přimělo o těch výrocích stále více přemýšlet se snahou dobrat se jejich hlubšího významu. Někdy jsem na dospělé útočil otázkami, které jim nebyly moc příjemné nebo je uváděly do rozpaků, takže mě různě odbývali s poukazem, že to časem sám pochopím. Nakonec z frustrace jsem se zhruba v období puberty sám proti sobě – příliš „dospělému“ – tak říkajíc vzbouřil a ke zděšení svého otce jsem se začal vyjadřovat jako běžný teenager. To se po jisté době zase změnilo, ale to už tady není důležité.

Proč zde tuto osobní historii vykládám, je myslím zřejmé. Nestačí, aby promluvy na sebe věrohodně navazovaly: člověk (v daném případě dokonce i probudilé dítě) má potřebu jejich skutečného *významu*, který má pro něho bytostnou závažnost. *Rozumět* druhému je právě schopnost přijímat a zase předávat takový význam. To myslím nelze očekávat od žádného automatu, jaké zatím známe: musel by mít *vědomí* – a pak by to už nebyl automat.

Vraťme se teď k původnímu tématu, které jsme nastolili na začátku. Špičkoví vědci v oboru umělé inteligence a robotiky odhalili ve svých prototypech umělých společníků člověka jeden závažný defekt: jejich roboti sice dokážou vyvolat iluzi poměrně inteligentní konverzace, ale chybí jim *empatie*, tedy schopnost vcítit se do citových stavů svého lidského protějšku. Člověk má afekty a emoce, také stálejší city nebo naopak okamžité nálady – a dokonalý robotický společník je musí umět rozpoznat a na ně adekvátně reagovat. Rozbor znaků, podle nichž se musí orientovat, zná každý zkušený psycholog, i když jej provádí většinou intuitivně: začíná výrazem tváře, tedy pohyby a stahy nebo uvolněním obočí, tkáně kolem očí, rtů a tváří kolem úst; pokračuje gesty rukou a celou *řečí těla*; neméně důležitá je i analýza jeho řeči, jak relativní výška nebo hloubka hlasu, kadence a způsob nasazení a doznění slov, ale samozřejmě také sama slova, která ten člověk zrovna používá a v nichž lze najít taková, jejichž význam má citové zabarvení, jejich četnost a podobně. Toto a mnoho dalšího dnes umělá inteligence začíná umět rozpoznat na základě rozborů, které předtím provedli její programátoři, ale také opět procesem učení se v tréninkových seancích. Není v tom ještě dokonalá, ale dá se očekávat, že i v tomto směru se bude stále zlepšovat.

Na zjištěný stav má ovšem robot *empaticky reagovat*, to značí odpovídat v souladném duchu, například vyjadřovat účast a podporu, respektive jemně povzbuzovat v rozpoznané rozladě či depresi. Co myslím přitom bylo málo (nebo nebylo vůbec) zdůrazněno, je možné selhání, když člověk vycítí *prázdnotu* či přímo *faleš* reakce. Jakkoliv emocionální stavy mohou logicky orientovaným vědcům připadat jako primitivní ve srovnání se sémantikou slov, problém je tady jemnější a v jistém smyslu zrádnější. Emoce jsou vývojově starší, patrně je mají i někteří jiní tvorové, kteří nemají řeč – a zřejmě proto naše citlivost na hodnověrné reakce je mnohem více vytříbená. Mohu se s někým věcně bavit o nějakém tématu – a když

mám pocit, že mi druhý v něčem špatně porozuměl, prostě mu to vysvětlím a je už třeba v tomto směru značné frustrace, abych tu konverzaci vzdal, protože zřejmě nemá cenu. Když naproti tomu budu mít například smutek a v útěšných slovech toho druhého zachytím byť jen slabou nepřiměřenost či faleš, nejspíše se od něho odvrátím a aspoň načas se mu uzavřu.

Předpokládejme ale, že i tato úskalí se postupným vývojem podaří překonat a že robot bude ve všech citových situacích reagovat bezchybně. Je ale něco takového? Jsou přece možné různé reakce. Vezměme případ bolesti nebo obecně utrpení, u něhož nejspíše je empatie žádoucí. Pokud půjde o živé lidi, kteří se mnou doopravdy cítí, budou jejich reakce často odlišné, podle jejich založení a také závažnosti bolesti: jeden či jedna bude brumlat soucitná nebo útěšná slova, druhý se třeba ze soucitu rozpláče, jiný naopak ztuhne a bude s bledou tváří bojovat se sdíleným utrpením a opět jiný, v kontrastu, rozvine překotnou aktivitu se snahou mi nějak pomoci. Tak by se dalo pokračovat dál, jistě by se našly i jiné hodnověrné reakce.

Může se ale také lišit má reakce na projevenou empatii: někdy ji budu vděčně přijímat, jindy naopak odmítat ať proto, že nechci dávat najevo slabost, nebo proto, že nechci blízkým přidělávat starosti, nebo i proto, že chci na svou bolest zapomenout nebo proti ní zmobilizovat své síly. Mohu mít pocit, že mě soucit druhých oslabuje nebo v určitém smyslu pokořuje, zpochybňuje mou roli nezlomného chlapíka.

Nicméně předpokládejme, že se dá najít optimální projev empatie pro daného člověka, případně že efektivní inteligence robota bude na takovém stupni, aby poznal, kdy jsou projevy soucítění žádoucí a kdy jsou naopak z různých důvodů odmítány. Zůstává otázka, zda robot se mnou doopravdy cítí, jakkoliv sugestivní dojem v tomto směru na mne udělá. Než na ni dáme svoji zkusmou odpověď, zdržme se ještě na okamžik v čistě lidském prostředí. Může opravdu člověk cítit utrpení druhého?

Ano, je ona bezprostřední reakce, kdy námi projede osten téměř opravdové bolesti, když vidíme, jak se někdo jiný zranil – o to se postarají ty už zmíněné *zrcadlové neurony*. Tak jako každý z nás má jistý *práh*, nad nímž už cítí bolest z fyzického podnětu, tak jsme i různě citliví v takové empatické reakci: jeden až může omdlít při pohledu na kapku krve někoho jiného, jiný stoicky bude přihlížet třeba i brutálnímu mučení. Když ale zůstaneme někde mezi těmito extrémy, tak empatie s utrpením druhého je vždycky nutně omezená. Když mě bude bolet například zub, abychom nezacházeli zas do extrému, nikdo jiný tu bolest doopravdy *fyzicky* cítit nemůže. Co ale může (a co se také běžně děje), je pociťovat *psychickou* trýzeň při pohledu na vnější příznaky mé bolesti.

Obdoba fyzického zranění či jiné bolesti v poněkud subtilnější formě existuje i v psychické oblasti. Mám někoho rád, což téměř nutně znamená, že si ho taky vážím, a jsem například svědkem toho, jak ho někdo ponížil nebo mu jinak způsobil psychickou újmu. Nemusím ani vidět, jak můj milý člověk zbledl nebo zrudl, jak se mu zkřivila tvář nebo mu slzy vstoupily do očí – pocítím psychickou bolest, někdy možná větší, než on sám, když si jasně uvědomím rozsah oné újmy; nicméně moje bolest není *jeho* a nejspíš nezpůsobí stejně silné trauma.

Naskýtá se nyní následující ošemetná otázka. Je dosti pravděpodobné, že se mě nepříjemně dotkne, když blízký člověk s mým trápením zjevně nemá žádnou empatii nebo ji sice dává najevo, ale já budu mít pocit, že ji jenom předstírá, že se mnou opravdu necítí. Ale

mohu si doopravdy *přát*, aby můj blízký člověk *trpěl*, byť to bylo proto, že mě něco bolí? Nejenom, že mi jeho utrpení přece nijak nepomůže (to je racionální úvaha), ale můj vztah k němu právě v sobě obsahuje přání, aby pokud možná nikdy netrpěl (to je citová stránka). Z toho důvodu také svoji bolest často skrýváme nebo alespoň předstíráme, že je slabší, než opravdu je. Přesto si paradoxně opravdovou empatii skoro vždycky přejeme, neboť nás jednak přesvědčuje o hloubce vztahu toho druhého, jednak je tak říkajíc záchranou z té velké osamělosti našeho trápení. Někde hluboko v nás je ukryto malé dítě, které se utíkalo k matce, aby utěšila jeho trápení – a tato potřeba se asi promítá do našeho paradoxního postoje.

Teď už se lze konečně ptát: může mi takovou empatii poskytnout sebedokonalejší robot? Pokud se zdaří všechno to, co si příslušní vědci předsevzali, bude mi asi schopen nabídnout její možná dokonalou iluzi. Platí však zde ještě víc to, co jsme už říkali o rozumění významu: *skutečně cítit* se mnou ovšem nemůže, protože *nemá cit*, má jenom schopnost dokonalé reakce. Tak tomu bude vždy, pokud bude jen strojem, který nemá vědomí. Jakmile si toto uvědomím, iluze se rozplyne.

Je možné, že v určitých stavech vědomí poněkud zastřeného utrpením nebo mozkovou degenerací iluze bude přesto přesvědčivá. Pak bude dosaženo cíle, který si vývoj empatických androidů vytyčil. Nejsem si ale jist etikou toho cíle. Smíme být takto klamáni právě tehdy, když jsme nejvíce bezbranní? Je dobré cokoliv, co efektivně zmírní moji osamělost v depresi či bolesti? To je otázka, na niž není lehká odpověď.