Inspirace pro umělou inteligenci (verze 0.31)

Adam Nohejl

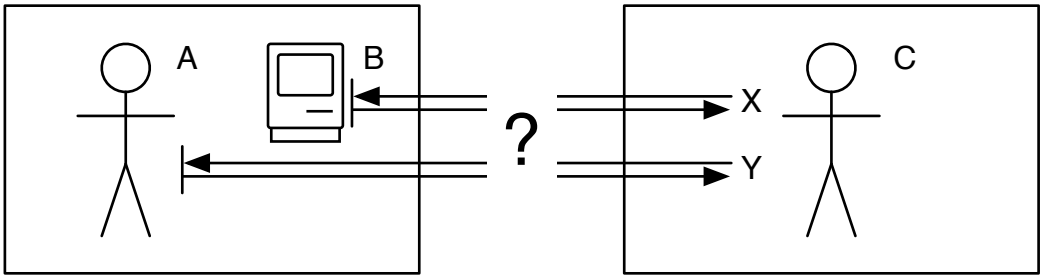
15. prosince 2010

První verze tohoto textu vznikla (stejně jako dříve *Uvažování o počtech, množstvích a číslech*) jako práce pro seminář Ivana M. Havla *Přirozené a umělé myšlení*. Teď prochází úpravami, aby se z něj časem mohl stát článek. Budu vám velmi vděčný za jakékoli komentáře. Časem se na stránce, ze které jste text stáhli, možná objeví aktualizované verze, případně podněty ke komentářům.

„Mohou stroje myslet?“ ptal se Alan Turing ve svém dnes klasickém článku [10] z roku 1950. Nespokojen s mnohoznačností takové otázky, přeformuloval ji pomocí následující hry pro tři hráče A, B a C:

- C je v místnosti oddělené od A a B a má s nimi dálnopisné spojení.
- C se na A a B obrací jako na X a Y a ti odpovídají na jeho otázky. Označení X a Y jsou od začátku hry daná, ale C neví, kterému z A a B přísluší které.
- Úkolem C je zjistit, který ze zbylých dvou hráčů je A a který B.
- Zatímco hráč A se snaží pomáhat hráči C, B se ho snaží zmást.

Ve variantě hry vhodné pro večírky se hráči A a B liší pohlavím. Jak však zřejmě tušíte, více nás zajímá případ, kdy hráč A je člověk a hráč B stroj. Tato verze (obrázek 1) popsané hry je dnes běžně známá jako *Turingův test* a stala se emblematickým experimentem umělé inteligence.



Obrázek 1: Turingův test.

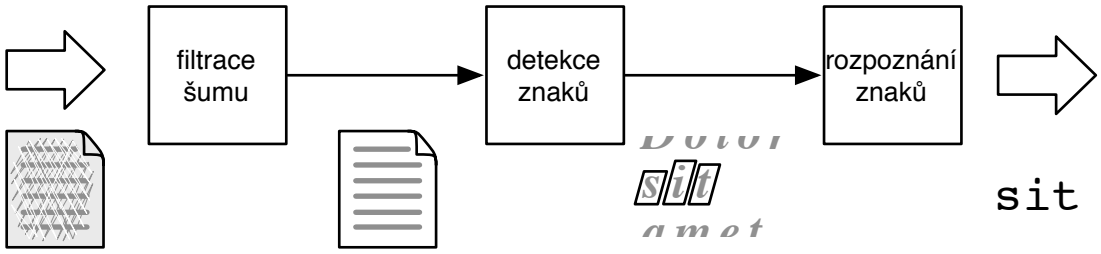
Turing byl stejně jako mnoho jeho současníku velmi optimistický: očekával, že stroj s programem, který ve většině případů projde pětiminutovým testem neodhalen, budeme mít k dispozici asi za padesát let. Zatím se tak nestalo a dřívější očekávání se z dnešní perspektivy jeví jako přehnaná nebo alespoň předčasná. Mimoto byl samotný Turingův test v uplynulém půlstoletí mnohokrát kritizován a většina dnešních vědců ho nepovažuje za zajímavé měřítko praktické úspěšnosti umělé inteligence.

Návrh testu a článek, který ho doprovázel, ale dobře ilustrují přístup, který je stále ve výzkumu umělé inteligence dominantní. Někdy bývá označován jako *komputacionalismus* a je založený na představě, že lidské myšlení se podobá výpočtu nebo lze alespoň výpočtem úspěšně imitovat. Alan Turing ostatně stál u zrodu teorie výpočtů, tedy základů počítačové vědy úzce spjatých s matematickou logikou, a na sklonku čtyřicátých let programoval jeden z prvních elektronických počítačů Manchester Mark 1. Tento počítač Univerzity v Manchesteru byl ve své době britským tiskem prezentován

příznačně jako „elektronický mozek“: tehdejší poznatky neurovědy se totiž zdály mít s principy elektronických počítačů mnoho společného a schopnost logického uvažování nebo provádění matematických výpočtů bychom u člověka jistě považovali za projev inteligence. Zaznívaly i opačné hlasy: proti označení „elektronický mozek“ se ostře ohradil Turingův kolega z katedry neurochirurgie, a Alan Turing si dobře uvědomoval, že jím navržený test neslouží ke zjištění, zda má stroj vědomí:

> Nerad bych vyvolal dojem, že si myslím, že na vědomí není nic záhadného. Je tu například určitý paradox spojený se všemi pokusy lokalizovat ho. Nemyslím ale, že tyto záhady musí být nutně vyřešeny, než budeme moci odpovědět otázku, která nás zajímá v tomto článku. [Ze závěru článku [10]. Stejně jako u dalších citací, můj překlad.]

Jinými slovy: otázku vědomí přenechme jiným a vytvořme stroj, který by imitoval ty aspekty lidské mysli, které nás zajímají. Vyjděme z toho, že mysl *reprezentuje* objekty reálného světa, jejich vlastnosti a vztahy pomocí *symbolů*; symboly pak zpracovává na základě *pravidel*, provádí tedy symbolický *výpočet*, na jehož základě vydá řešení problému nebo jinak jedná. Tato teze může obstát i v jiných vědách: souzní s východisky neurovědy a v psychologii ji můžeme označit jako *kognitivismus*[1]. Jejímu udržení ve výzkumu umělé inteligence napomohl i plynulý přechod mezi praktickým řešením specifických úloh pomocí běžných programátorských technik a algoritmů, tedy denním chlebem počítačových inženýrů a vědců, a dlouhodobým cílem vytvoření obecné umělé inteligence založené na principech reprezentace a výpočtu. Pro běžného programátora reprezentaci pomocí symbolů odpovídají *datové struktury* a pravidla pro výpočet jsou dána *algoritmem*, v pojmech komputacionalismu se tedy cítí jako doma.



Obrázek 2: Počítačové rozpoznávání tištěného textu (zjednodušený postup).

V průběhu druhé poloviny dvacátého století komputacionalistický přístup dosáhl významných úspěchů při řešení dílčích problémů. Například z vnímání lze vymezit rozpoznávání, a z už tak úzkého oboru vybrat úlohy rozpoznávání lidských tváří, rozpoznávání tištěného textu nebo rozpoznávání řeči. Všechny tyto problémy umíme dnes na počítači velmi dobře řešit a zpravidla se tak děje jejich rozdělením na ještě menší podproblémy, na něž lze aplikovat velmi specializovaný algoritmus: rozpoznávání tištěného textu může například sestávat z odfiltrování šumu, detekce řádků, rozdělení na znaky, odhadu sklonu písma, rozpoznání znaků a opravy chyb pomocí slovníku (schematicky viz obrázek 2, skutečný seznam by byl ještě delší a v některých bodech méně intuitivní). Prozatím však selhávají pokusy o obecnou umělou inteligenci. Posledním takovým velkým pokusem byl projekt tzv. páté generace počítačů v Japonsku zahájený v osmdesátých letech. Jeho hlavním cílem bylo vytvořit systém založený na logickém odvozování, který by rozuměl lidskému jazyku a byl schopný sám pro sebe psát programy podle zadání neškoleného uživatele. Tento cíl se nepodařilo uskutečnit.

Současně začaly přicházet i nové, nebo znovuobjevené, přístupy ke strojovému řešení úloh. Nejviditelnější skupinu, která si už vydobyla určité místo v hlavním proudu, jsou *konekcionistické* modely

[1] Toto označení používá v kontextu umělé inteligence, neurovědy a především filosofie mysli, resp. kognitivní vědy, Francisco Varela a jeho spolupracovníci v knize *The Embodied Mind* [11]; jejich shrnutí vlastností kognitivismu (ve 3. kapitole knihy) přejímám. V rámci umělé inteligence používám pro totéž běžnější označení *komputacionalismus*, a to s vědomím, že jde spíš o umělou nálepku pro přístup, který je v oboru dosud převažující.

založené na *emergentních* vlastnostech. Většinou se jedná o různé druhy umělých *neuronových sítí*, které ovšem mohou mít k živým neuronům velmi daleko. Spojuje je s nimi často jen premisa, že zajímavé složité chování celku vzniká (proto emergence) spojením (proto konekcionismus) mnoha stavebních prvků, jejichž chování je definované jednoduchými pravidly.

Protože v reálných aplikacích platí, že *cokoli funguje, je správně*, je dnes běžné konekcionistické a komputacionalistické přístupy podle potřeby kombinovat. V úlohách, které jsem uváděl jako příklady úspěchů komputacionalistického přístupu, se často jako jeden z článků zpracování používá nějaký druh neuronové sítě (například pro klasifikaci předzpracovaných znaků při rozpoznávání textu).

Další, možná ještě různorodější skupinou, jsou různé metody inspirované biologií: může jít o *evoluční algoritmy*, inspirované darwinovskou evolucí, nebo *inteligenci hejna*, založenou na chování skupin zvířat. Lze sem zahrnout i už zmíněné neuronové sítě. Inspirace biologií může nabývat mnoha svérázných podob a zpravidla jde o inspiraci velmi volnou, vybíravou, opět podle hesla *cokoli funguje, je správně*. Výsledný model může a nemusí mít konekcionistické nebo emergentní vlastnosti. Mnoho z těchto přístupů se dnes používá pro praktické řešení problémů, jež by klasickými algoritmy byly obtížně zvladatelné.

Paralelně s vývojem umělé inteligence začala vznikat nová disciplína, která v sedmdesátých letech nabyla jasných obrysů a označení *kognitivní věda*. Kognitivní věda je je mezioborové studium spojující psychologii, neurovědu, biologii, informatiku, jazykovědu, filosofii a další obory dotýkající se lidské nebo umělé mysli. Od počátku byla orientovaná spíš prakticky a v duchu komputacionalismu, což je stále převažující náhled: v populární internetové encyklopedii [12] se můžeme například dočíst, že kognitivní věda se zabývá tím, „jak je informace *reprezentována* a *transformována* mozkem“ (moje zvýraznění): představa z vnějšku vstupující informace, její reprezentace a následného zpracování podle pravidel, a předpoklad, že mysl lze redukovat na mozek, jsou zřejmé.

V devadesátých letech se v kognitivní vědě objevil nový náhled na lidskou mysl, čerpající především z fenomenologie a stavějící na *vtělené* (angl. *embodied*) mysli a principu *zjednávání* (angl. *enaction*)[2]. Jako jedni z prvních tento přístup rozvinuli vědci vedení Franciskem Varelou v knize *The Embodied Mind: Cognitive Science and Human Experience* [11]. Enaktivismus nevychází z počítačové vědy, umělé inteligence ani analytické filosofie ideově blízké komputacionalismu. Nenabízí tedy okamžité praktické uplatnění v umělé inteligenci. Pro nedostatek fungujících aplikací, by tedy bez ohledu na jeho přínos pro kognitivní vědu jako celek, bylo předčasné tvrdit, že enaktivismus je správný, protože funguje. Z perspektivy počítačového vědce, který by připustil, že takový přístup má v jeho praxi uplatnění, by se enaktivismus mohl jevit jako další z řady méně ortodoxních přístupů (neuronové sítě, evoluční algoritmy, inteligence hejna), který ovšem přichází bez jasného návodu k použití.

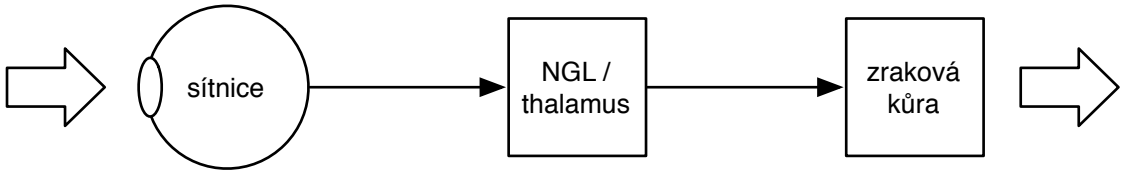
Pojďme se podívat, jak pohled na řešení některých problémů změnily metody založené na emergenci a jedna z biologií inspirovaných metod, evoluční algoritmy. Uvidíme pak jasně, v čem pro umělou inteligenci spočívá přínos enaktivního přístupu a přijetí teze vtělené mysli a co přesně se za tímto náhledem skrývá.

Vzestup emergence

Když jsme mluvili o komputacionalismu, narážel jsem na styčné plochy s určitým pohledem na lidský mozek. Vynecháme-li povrchní podobnost ve využití elektrického signálu jako přenosového média, mluví se často i v souvislosti s lidským mozkem o reprezentaci a vysokoúrovňové myšlenkové procesy bývají viděny jako logické odvozování, vykonávání složitých postupů (algoritmů) nebo provádění náročných výpočtů. Reprezentaci lze spatřovat v jakémkoli vzorci mozkové aktivity: neurovědci skutečně umí vysledovat korelaci mezi mozkovou aktivitou v určitých oblastech a konkrétními vjemy nebo myšlenkami, z čehož vzniká populární představa, že dostatečně přesným mapováním

[2] *Zjednávání* jako český překlad *enaction* používá Ivan M. Havel, viz např. jeho článek *Zjednaný svět* [5].

mozkové aktivity by bylo možné číst myšlenky, ale také vážně míněné diagramy, kde například „zpracování“ vjemu ze sítnice vypadá jako sekvenční zpracování signálu sérií skříněk, na jehož konci je odpovídající reprezentace (obrázek 3). Nepřipomíná vám to náš krátký příklad rozpoznávání textu počítačem?



Obrázek 3: Sériové „zpracování signálu“ zrakového vjemu. Dnes už překonaná představa, i když v učebnicích se stále objevuje. Upraveno z [11].

Na této úrovni popisu je dobře patrná podobnost se symbolickými výpočty. Pod ní se ovšem skrývá mnoho jednoduchých jednotek, neuronů, provázaných ještě větším množstvím spojům, které pracují paralelně a distribuovaně (tedy ne sekvenčně). Při pohledu na úroveň neuronů není žádná organizace do škatulek vidět, jednotlivé neurony navíc fungují všechny velmi podobně podle jednoduchých pravidel. Díky vhodným pravidlům na úrovni neuronů tedy může dojít k samoorganizaci na vyšší úrovni.

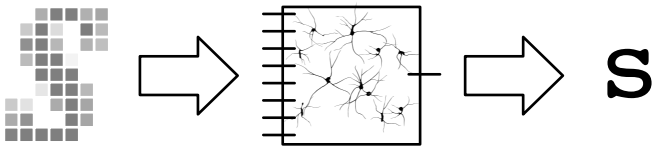
V živé neuronové síti i v počítačovém modelu předpokládáme, že pokud předložíme neuronové síti (prostřednictvím změny elektrických potenciálů na nervových zakončeních, resp. změny hodnot na umělých neuronech ve vstupní vrstvě) vícekrát za sebou nějaký *vzor*, dojde k jeho *naučení*, tedy ustálení nějakého konkrétního stavu celého systému, a neuronová síť tak vzor bude schopna odlišit od jiných, *klasifikovat* ho. Zřejmě nejznámější takové učicí pravidlo formuloval neuropsycholog Donald Hebb v roce 1949. Bývá parafrázováno „cells that fire together, wire together“ a stojí za ním idea, že mimo rychle se měnících elektrických potenciálů se v neuronové síti pomalu mění síla jednotlivých spojů, a to tak, že spoje buněk, které bývají současně aktivní, se posilují.

První modely inspirované neurony se objevovaly ve čtyřicátých až šedesátých letech. Krátce na to byly zavrženy a pozornost se k nim vrátila opět až v osmdesátých letech. Takové umělé neuronové sítě mohou mít různou architekturu, různá pravidla pro excitaci neuronů a různé postupy pro učení. Například Hopfieldovu asociativní síť navrženou v roce 1982, kterou bychom mohli použít pro rozpoznávání písmen v naší úloze rozpoznávání textu, lze učit postupem odvozeným od Hebbova pravidla. Ve skutečnosti je výhodnější použít jiný algoritmus, variantu tzv. perceptronového učení, o němž lze matematicky dokázat, že je v jistém smyslu ekvivalentní a přitom robustnější. Opět je přijatelné, cokoli funguje. Inspirace hypotézou z neuropsychologie může být užitečná, ale pokud se nám nehodí, odhlédněme od ní: pro dosažení jiných zajímavých vlastností lze Hopfieldovy sítě kombinovat například se *simulovaným žíháním*, optimalizační metodou inspirovanou žíháním v metalurgii. V sofistikovanějších modelech neuronových sítí pak zpravidla najdeme více aplikované lineární algebry a kalkulu než inspirace neurovědou. Užitečné je také dokazovat vlastnosti modelu z pohledu teorie výpočtů, resp. výpočetní složitosti.

Co nám tedy konekcionistické modely přinesly? Samoorganizaci, učení, adaptaci. Chceme-li rozpoznávat písmena, nemusíme zjišťovat, z čeho se skládají nebo jak se od sebe navzájem liší; nemusíme se starat o to, jakým přesně písmem jsou vysázená nebo zda mohou být špatně vytištěná: s tím vším se vhodně zvolená neuronová síť sama vypořádá. Nejde samozřejmě o všelék, slova „vhodně zvolená“ jsou v tomto případě klíčová: konkrétní model (typ sítě, učicí algoritmus) s danými parametry (například počet jednotek a struktura spojů), který funguje pro jednu úlohu, neobstojí v jiné, a pro některé problémy se dosud známé typy neuronových sítí nehodí vůbec.

Zjevný rozdíl pak najdeme ve fungování systému: kognitivním procesem je tu vznik, tedy emergence, jeho globálních stavů, nikoli reprezentace pomocí symbolů a jejich výpočetní zpracování. Přesto je možné komputacionalismus a konekcionismus vidět jako doplňující se přístupy: *symbolický*

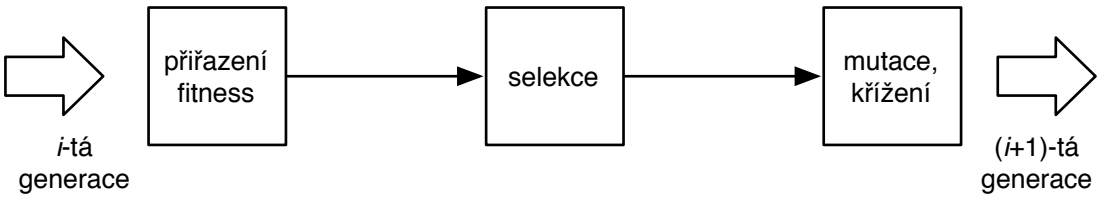
a *subsymbolický*: globální stavy konekcionistického modelu můžeme vidět jako reprezentace nebo symboly výpočetního zpracování. To se také přímočaře přenáší do praxe: zapojíme-li neuronovou síť do našeho systému pro rozpoznávání textu, tak systém vloží na vstup neuronové sítě obraz písmene reprezentovaný bitovou mapou a po jeho zpracování zjistí stav neuronové sítě, se kterým dále pracuje jako se symbolem pro písmeno. Od našeho původního postoje jsme se nemuseli odchýlit příliš. Neuronovou síť můžeme použít jako černou skříňku se vstupem a výstupem v našem dosavadním modelu: kafemlejnek, do kterého z jedné strany nasypeme naši reprezentaci dat, a když zatočíme klikou, tak z druhé strany vypadne symbol (obrázek 4).



Obrázek 4: Použití neuronové sítě jako kafemlejnku pro řešení dílčí úlohy: reprezentace rozpoznávaného písmene maticí hodnot na vstupu, symbol na výstupu. (Grafika neuronů z písma Neurons, © 2009, Joyce LI Yan Yee.)

Evoluční algoritmy: přirozený nebo umělý výběr?

V šedesátých letech se začaly objevovat první pokusy o využití principů darwinovské evoluce pro řešení výpočetních problémů. Společná idea evolučních výpočetních metod zpočátku spočívala v náhodné aplikaci mutace na reprezentace řešení úlohy a postupném nahrazování méně způsobilých řešení způsobilejšími přirozeným výběrem.[3] Dále byly tyto ideje rozvinuty v sedmdesátých letech Johnem Hollandem [6] jako *genetické algoritmy*: mimo mutace se začalo používat křížení, proces probíhal na celé populaci řešení, byly formulovány teoretické výsledky týkající se selekce na základě dosažené hodnoty *fitness*. V devadesátých letech byla navržena obdobná metoda tentokrát uzpůsobená přímo pro vývoj počítačových programů: *genetické programování*. Jmenované postupy a mnoho dalších od nich odvozených se souborně nazývají *evoluční algoritmy*.



Obrázek 5: Evoluční algoritmus: populace jedinců (kandidátních řešení) jako objekt formujících tlaků vnějšího prostředí, tedy funkce *fitness*, selekce, mutace a křížení.

Podívejme se tedy podrobněji na to, co jsme zatím jen načrtli (obrázek 5): Jak je možné provádět mutaci nebo křížení na nějaké počítačové reprezentací řešení úlohy? V prvé řadě musíme nějakou reprezentaci navrhnout (systém si ji sám vytvořit neumí), a pak odpovídajícím způsobem naprogramovat mutaci a křížení: tak, aby dobře fungovaly v naší úloze a se zvolenou reprezentací. Mutaci a křížení se v kontextu evolučních algoritmů říká *operátory*. Opět tedy vycházíme z komputacionalistického pohledu: máme reprezentace, na kterých provádíme operace. Mohl by tu však být podstatný rozdíl: hybnou silou evoluce má být přeci přirozený výběr. V evolučních algoritmech se zpravidla

[3] V češtině se vžilo překládat anglický *survival of the fittest*, pojem už tak mající zavádějící konotace, jako „přežití silnějšího“. *Fit* přitom neznamená silný, ale vhodný, schopný, kompetentní, způsobilý. Tento rozpor vynikne u pojmu *fitness*, který lze překládat jako „způsobilost“ nebo „zdatnost“, rozhodně ne „síla“. V následujícím textu používám raději anglické *fitness* také proto, že v kontextu evolučních algoritmů se termín zpravidla nepřekládá.

mlčky vychází z danosti vnějšího prostředí, které vlastně přirozený výběr jedinců provádí. Jde se dokonce o krok dál a nepředpokládá se ani žádná forma přímé soutěže mezi jedinci při rozmnožování. *Fitness* je místo toho jedincům deterministicky přiřazena, jsou ohodnoceni tzv. funkcí *fitness*, která zastává roli onoho vnějšího prostředí, a výběr pro reprodukci se provede poměrným zastoupením podle dosažených hodnot. To je ovšem darwinovský výběr naruby: nejdříve má probíhat interakce jedince s prostředím (které prozatím považujeme za dané a oddělené od jedince), což zahrnuje i soutěž s ostatními jedinci a samotnou reprodukci, až posléze z toho můžeme my jako pozorovatelé usuzovat o jeho *fitness*.

Jedním z pojmů klíčových pro evoluci, který vzešel z tzv. *moderní syntézy* darwinismu, mendelovské genetiky a molekulární biologie, je *rozlišení genotypu a fenotypu*. Genotyp je děděná genetická informace, zatímco fenotyp je soubor charakteristik pozorovatelných při interakci organismu s prostředím. Triviální chápání tohoto rozdílu vede k představě *mapování genotypu na fenotyp* ve smyslu matematické funkce, zobrazení. Pokud se vůbec oddělení fenotypu od genotypu v evolučních algoritmech uvažuje, bývá pojato v tomto duchu a považuje se za žádoucí, aby mapování mělo různé pravidelné vlastnosti, aby se chovalo předvídatelně. Rádi bychom například, aby malá mutace genotypu vedla ke změně malého počtu charakteristik.

Tento směr úvah vyústil při zrodu genetických algoritmů ve *větu o schématech* (formuloval ji Holland [6]) a v úzce související *hypotézu stavebnicových kostek* (angl. *the building block hypothesis*, pod tímto názvem ji zavedl David Goldberg [4]). Věta o schématech rigorózně popisuje za jakých podmínek genetický algoritmus konverguje k optimálnímu řešení nějakého problému. Abychom ji mohli aplikovat v tomto smyslu konvergence k optimu, potřebujeme splnit mimo jiné dva předpoklady:

Zaprvé je třeba zbavit se výběrové chyby, tedy chyby způsobené výběrem statisticky nereprezentativního vzorku. Výběrovou chybu lze minimalizovat volbou vhodných postupů, přesto však může být pro netriviální problémy velmi vysoká. Protože genetický algoritmus pracuje náhodně na relativně malých populacích, chyba se v něm nutně projeví. Principiálně stejné výběrové chyby nastávají i u živých populací, čímž vzniká tzv. *genetický drift*.

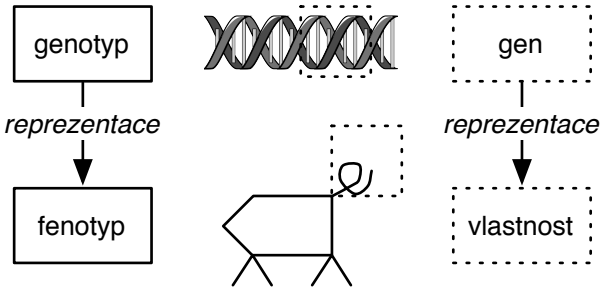
Zadruhé musí platit hypotéza stavebnicových kostek, která tvrdí, že skoro optimální řešení se skládá z parciálních řešení s nadprůměrnou *fitness*. Jinými slovy: *fitness* lze zvyšovat malými změnami a jejich prosté složení dává skoro optimální řešení. Goldberg ve své knize [4] hypotézu podpíral „rostoucím souborem empirických důkazů [. . .] z různých tříd úloh“, které „napovídají, že genetické algoritmy jsou vhodné pro mnoho typů problémů, s nimiž se běžně setkáváme“. Na otevřenosti hypotézy se od té doby ovšem nic nezměnilo, naopak se potvrzuje, že se běžně setkáváme jak s případy, kdy evidentně platí, tak s případy, kdy se nám ony stavební bloky nedaří nalézt. Zřejmě by bylo vhodnější mluvit o vlastnosti mapování genotypu na fenotyp pro určitou úlohu, ne o obecné hypotéze. Pokud problém a zvolené mapování tuto vlastnost nemá, jsou naopak fenotypické efekty změn genomu navzájem složitě provázané. I to má svou biologickou obdobu, tzv. *pleiotropii*.

Od osmdesátých let došlo k vývoji nových evolučních algoritmů i jejich teoretického zázemí. Trhlina, která zůstala po původní nadějné teorii kolem genetických algoritmů zůstává ale otevřená. V praxi se výše načrtnuté obecné potíže řeší, pokud to vůbec jde, případ od případu a bez ohledu na ideovou čistotu. Jestliže jsme tedy shledali, že genetické algoritmy sdílejí principy neodarwinismu jen částečně, v praxi je běžné odchýlit se od nich podstatně víc.

Je pozoruhodné, že do velké míry je tento vývoj paralelní[4] s vývojem v evoluční biologii: dva zmíněné jevy, genetický drift a pleiotropie, bývají často používány jako argumenty proti ortodoxnímu adaptacionismu. Přísní neodarwinisté se snaží význam těchto jevů naopak bagatelizovat. Známý autor *Sobeckého genu* Richard Dawkins tvrdí, když svou teorii rozvádí v *The Extended Phenotype* [2], že genetický drift není „slabina přirozeného výběru“, ale „může teoreticky zvýšit pravděpodobnost, že vývojová linie dosáhne optimálního designu“, „stejně jako v případě pleiotropie tu není žádný paradox.“. Dawkins pak *replikátory* (ony populární sobecké geny) definuje v 5. kapitole [2] obdobně

[4] Slovo *paralelní* je skutečně namístě. Dawkinsovy knihy *Selfish Gene* a *Extended Phenotype* [2] byly vydané v letech 1976 a 1982, Hollandova kniha [6] v roce 1975 a Goldbergova [4] v roce 1989.

jako Goldberg své stavebnicové kostky a Holland schémata odpovídající předpokladům věty o schématech: jde o krátké fragmenty genomu, které mají určitý samostatný význam a replikují se pro svou vysokou *fitness*. Přirozený výběr je pak primárně viděn na úrovni replikátorů a evoluce organismu je složením evoluce těchto základních jednotek. S adaptacionistickým viděním evoluce má pak původní teorie genetických algoritmů společné i zjednodušené chápaní mapování genotypu na fenotyp a snahu o dělení světa na vyvíjející se jedince a fixní prostředí.



Obrázek 6: Mapování genotypu na fenotyp jako prostá reprezentace: na úrovni jedince a na úrovni „stavebních kostek“ jeho genomu.

Aniž bych chtěl napadat legitimitu ortodoxního darwinismu, adaptacionismu nebo teorie „sobeckých genů“ v rámci jejich oboru, musím konstatovat, že do počítačové vědy tato linie myšlení nepřinesla fundamentálně nové myšlenky. Odpovídající úvahy nás v případě genetických algoritmů vedly vlastně k zavedení symbolů a reprezentace na další úrovni: stavebnicové kostky reprezentují jednotlivé vlastnosti, jsou to (sobecké) geny pro tyto vlastnosti (obrázek 6).

Na závěr bych rád vyzdvihl zjevný přínos evolučních algoritmů v několika oblastech: ze své povahy jsou schopny už v průběhu práce dodávat alespoň nedokonalá řešení úlohy. Nabízejí se různé možnosti jejich paralelizace. Idea odděleného genotypu a fenotypu, aplikovaná v některých typech evolučních algoritmů, může být dále rozvíjena i jiným směrem než ke „stavebnicovým kostkám“.[5] Všechny tyto body lze chápat jako odchylky od toho, co je typické pro komputacionalistický přístup, i když jeho jádro nenapadají. Všimněme si však, že žádný z bodů není závislý na principech adaptacionismu.

Enaktivní přístup: jednání první

Francisco J. Varela, Evan Thompson a Eleanor Rosch v knize *The Embodied Mind: Cognitive Science and Human Experience* [11] předestírají diskusi dosavadních přístupů v kognitivní vědě a souvisejících oborech a souběžně rozvíjejí vlastní nový přístup ke zkoumání mysli, který staví na myšlenkách fenomenologických filosofů, především Maurice Merleau-Pontyho, a buddhistické meditační tradici. V centru jejich zájmu tedy není umělá inteligence, ale kognitivní věda. Pokusím se tedy jejich přístup představit alespoň částečně v tomto rámci.

Jedno z východisek autorů je hledání *já* v západní filosofii, které jim slouží jako odrazový můstek pro spojení s filosofií východní. S odkazem na Huma poukazují na to, že seriózní pokusy o nalezení nějakého pevného *já* skončily neúspěchem, který ale nejsme schopni nějak dále reflektovat nebo rozvinout. Buď rozpor s naší každodenní zkušeností, která *já* zahrnuje, ignorujeme nebo podobně jako Kant předpokládáme transcendentální ego. V tomto okamžiku pouštějí ke slovu buddhistickou meditační tradici. Jak autoři zdůrazňují, meditace není soustředění nebo uvolnění, trans nebo mystické prožření, jak toto slovo chápeme zvulgarizovaně, ale především technika sloužící ke zkoumání naší zkušenosti a vědomí. Může tak tvořit doplněk k fenomenologii, která přijímá introspekci jako platný prostředek poznání.

[5] S takovým netriviálním mapováním genotypu na fenotyp experimentují například Michael O'Neill a Conor Ryan v knize *Grammatical Evolution* [9].

Pomocí buddhistického učení Abhidharmy a analýzy mysli v tzv. pěti agregátech autoři ukazují, jak východní tradice také dospívá k tomu, že naší zkušenosti žádné skutečné *já* přístupné není. Toto zjištění je paralelní s myšlenkami Descartesa a především Kanta. Dále se s nimi ale rozchází: Jaký může mít vztah Kantovo čisté, původní a neměnné *já*, k námi cítěnému *já*, k naší zkušenosti? Jak může být základem naší zkušenosti a přitom jí být nedotčeno? Meditační tradice vychází ze zkušenosti, která je okamžiková (sestávající z okamžiků, nespojitá) a *já* v ní v každém momentu vzniká jako důsledek našeho lpění, naší snahy uchopit se/*já*, (angl. *clinging*, *grasping to an ego-self*). Buddhismus toto zjištění dále rozvíjí: lpění na *já* je zdrojem našeho utrpení; meditační praxe má pak přispět k získání přímé hluboce přeměňující zkušenosti s touto prázdnotou *já*. Poskytuje tak to, co v naší západní kultuře založené na vědeckém poznání chybí:

> V naší kultuře věda přispěla k probuzení tohoto vědomí absence pevného *já* ale popsala ji jen zdálky. Věda nám ukázala, že pevné *já* není nezbytné pro mysl, ale neposkytla žádný prostředek, kterým bychom se mohli vypořádat s tím, že právě na tomto už nepotřebném *já* všichni lpíme a je nám tak drahé. [11]

Druhá, paralelní linie jejich úvah se týká vlastních kognitivních procesů: Představují kognitivistickou hypotézu, o níž jsme zde mluvili jen v kontextu umělé inteligence jako o komputacionalismu, a s ní spjatý reprezentacionismus. Shledávají, že přijetí nebo využití emergentních vlastností a prvků nevede k upuštění od kognitivismu, což jsem zde demonstroval na konkrétních příkladech z umělé inteligence. Věnují se i souvislostem kognice s evolucí (ne s evolučními algoritmy) a dochází mimo jiné k tomuto závěru: „Řečeno bez obalu, reprezentacionismus v kognitivní vědě je naprosto homologický k adaptacionismu v teorii evoluce, protože optimalita hraje stejnou roli v obou případech." Jinou cestou, přes pojem optimality, tedy zjišťují o přístupech v kognitivní vědě a evolučních teoriích totéž, co jsem se snažil ukázat na jejich aplikacích v umělé inteligenci.

Jako alternativu ke kognitivismu navrhují vycházet z myšlenek kontinentálních filosofů: hermeneutiky Heideggera a Gadamera a především z rané práce Merleau-Pontyho, která se dotýkala vědeckého výzkumu kognice (*Phénoménologie de la perception*, 1945). Z hermeneutiky si berou závislost naší mysli na našem vtělení, tedy „bytí ve světě neoddělitelném od našich těl, našeho jazyka, společenské historie" (odtud *embodied mind*). Od Merleau-Pontyho přejímají koncept kognice jako vtěleného jednání a na tomto základě definují *zjednávání* (angl. *enaction*): „(1) vnímání spočívá v perceptuálně vedeném jednání a (2) kognitivní struktury vznikají z opakujících se sensomotorických vzorů, které dovolují, aby jednání bylo vedeno perceptuálně." Vnímání, na rozdíl od kognitivistického pojetí, není zpracování informací z předem daného prostředí; prostředí se totiž neustále mění vlivem jednání vnímajícího, je tímto jednáním vlastně spoluvytvářeno. Nezajímá nás nezávislý vnější svět, ale situace toho, kdo vnímá, způsob, jakým je do světa vtělen. Nejde o získávání věrné reprezentace světa: svět je totiž závislý i na pozorovateli.

Enaktivní přístup v praxi

Krátce jsme načrtli, jaký je význam enaktivismu v kognitivní vědě. Jedná se o komplexní přístup, a není těžké si představit, jak lze s jeho pomocí přistupovat k otázkám souvisejících oborů, například evoluční biologie nebo psychologie, kde se přímo uplatňuje navržené doplnění introspektivními technikami převzatými z východní tradice.

Jak ale předestřené principy aplikovat ve výzkumu umělé inteligence? Varela, Thompson a Roschová se odvolávají na už probíhající výzkum v robotice, konkrétně na výsledky laboratoře umělé inteligence MIT vedené Rodney Brooksem. Ten ve svém článku *Intelligence without representation* [1] píše:

> [...] měli bychom vytvořit úplné inteligentní systémy, které vypustíme do skutečného světa se skutečnými vjemy a skutečnými akcemi. Cokoli menšího by nám umožňovalo klamat sebe sama. S tímto přístupem jsme postavili sérii samostatných mobilních robotů.

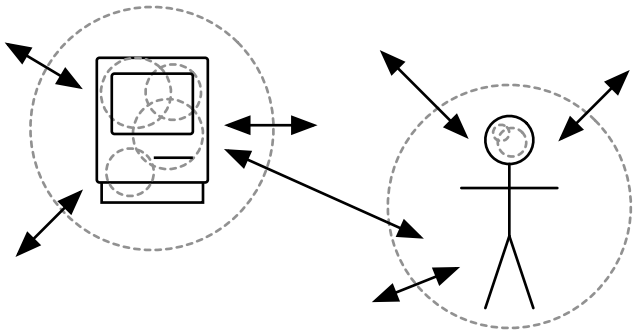
Došli jsme k nečekanému závěru (Z) a máme poměrně radikální hypotézu (H).

(Z) Při zkoumání velmi jednoduchých úrovní inteligence zjišťujeme, že explicitní reprezentace a modely světa jsou prostě jen na překážku. Ukazuje se, že je lepší používat svět jako model sebe sama.

(H) Reprezentace je špatná úroveň abstrakce pro budování hlavních částí inteligentních systémů.

Brooksem navrhované inteligentní systémy jsou potom místo na principu abstrakce založené na principu vtělenosti. Přestože se skládají z částí a vrstev, části nemají přesně vymezené funkce (jako je vnímání nebo jednání), a vrstvy jsou na sobě nezávislé. Zajímavé je, že jak Brooks sám uvádí, jeho přístup „se do jisté míry podobá pracím inspirovaným [Heideggerovou filosofií], ale [jeho] práce jím inspirována nebyla. Je založená výhradně na inženýrských úvahách."

Enaktivní nebo vtělený přístup má tedy, stejně jako dříve zmíněné metody, aplikaci, v níž prakticky funguje. Není toto praktické využití ale omezené třeba právě jen na jednoduchou mobilní robotiku? To je oprávněná námitka, kterou vznáší Daniel Dennett ve své velmi kritické recenzi [3] knihy *The Embodied Mind*, a dále tvrdí: „Problém je, že jakmile se pokusíme rozšířit Brooksovo zajímavé a důležité sdělení za hranice jednoduchých tvorečků (umělých nebo biologických), můžeme si být docela jisti, že něco *strašně podobného* reprezentacím se do toho *nutně* vkrade jako příliv, ve velkých vlnách." Pro toto své přesvědčení ale neuvádí jediný argument, jedná se tedy spíš o výraz vysoké nedůvěry, který tento přístup vzbuzuje u vědců a myslitelů držících se osvědčeného hlavního proudu.



Obrázek 7: Inteligentní systém vtělený do světa, zjednávající si svět. Uvnitř má „síť sestávající z více úrovní propojených, sensomotorických podsítí" (podle [11]). Slovo *sensomotorický* zde můžeme chápat i přeneseně: nejen ve významu fyzického pohybu, podstatné je, že vnímání je spjato s jednáním.

Podle mě je vtělenost mysli a zjednávání třeba chápat jako obecné principy, které neznamenají, že inteligentní systém musí mít fyzické tělo, nebo dokonce tělo podobné lidskému, ani že si musí zjednávat svět podobný našemu. To by byly zřejmě nutné podmínky pro vytvoření humanoidního robota, o to primárně nejde Brooksovi, Varelovi ani většině vědců pracujících ve výzkumu umělé inteligence. Konkrétní způsob vtělení samozřejmě vymezuje možnosti systému, nic však nebrání tomu, aby se tyto principy odrazily třeba i u čistě softwarového systému (obrázek 7). I ten je zasazen do nějakého světa: buď jako kafemlejnek transformující vstup na výstup, nebo tak, aby si své prostředí mohl zjednávat.[6]

V první kapitole *The Embodied Mind* najdeme ještě diagram příslušnosti vědců z různých oblastí ke kognitivismu, emergenci a enaktivnímu přístupu. Jako zastánce enaktivního přístupu v umělé

[6] To okamžitě evokuje dva typy uživatelského rozhraní, zpravidla označované dávkový a interaktivní. Uvědomme si ovšem, že uživatelské rozhraní je jen jedno z mnoha rozhraní, kterým software disponuje na své nejvyšší úrovni. Přes zřejmou paralelu nelze návrh softwaru nebo inteligentních systémů redukovat na návrh uživatelského rozhraní ani naopak. Poznamenejme ještě, že i grafická uživatelská rozhraní se často bohužel chovají dávkově, a také že přínosnější je v případě uživatelského rozhraní obrátit perspektivu, které jsme dosud užívali: zajímá nás totiž, jakým způsobem si svět může zjednat uživatel. Vynikající kniha na toto téma je *The Design of Everyday Things* Donalda A. Normana [8].

inteligenci je v diagramu umístěn vedle Brookse bez dalšího vysvětlení v textu John Holland, o kterém jsem mluvil v souvislosti s genetickými algoritmy. Spolu s Dennettem [3] v tomto případě musím vyjádřit údiv. Není jasné, které prvky Hollandova výzkumu zahrnují vtělenost nebo enaktivismus, jistě to nejsou genetické algoritmy, kterými se proslavil. I přesto, že Holland nebo Goldberg enaktivní přístup nezvolili, jsem přesvědčený, že oblast evolučních algoritmů se k jeho aplikaci nabízí.

Dennett ve své recenzi také poznamenává: „[autoři] při vysvětlování klíčového pojmu, ‚enaktivní', silně čerpají z tvrzení genetika Richarda Lewontina, že evoluci je třeba chápat z enaktivní perspektivy. [. . .] Lewontin typicky trvá na tom, že organismus hraje významnou roli v utváření svého vlastního prostředí." Nevím o tom, že by Lewontin sám používal slovo *enaktivní*, ale blízkost Varelova pohledu na vztah organismu a prostředí s Lewontinovým je zjevná; Varela to, alespoň v kapitole týkající se evoluce, dokládá jeho častými citacemi. I některé závěry v jiných kapitolách jsou obdobné jako v knize Lewontina, Rose a Kamina *Not in Our Genes* [7], která se zabývá evoluční biologií a genetikou v kontextu společenských věd a psychologie. Na rozdíl od Dennetta, který s Lewontinovými názory nesouhlasí, to nepovažuji za handicap. Naopak v tom vidím potvrzení, že enaktivní přístup je skutečně aplikovatelný i v oborech navazujících na kognitivní vědu.

Viděli jsme, že evoluční algoritmy přispěly k rozvoji umělé inteligence v několika směrech, z nichž ani jeden není závislý na tezích adaptacionismu. Mohli bychom tedy tedy některé jejich adaptacionistické, resp. komputacionalistické, principy („zpracování" jedinců prostředím, snaha o optimální „stavebnicovou" reprezentaci) nahradit enaktivními. To se pouštím na pole čiré spekulace: je možné, že v praxi by se tato myšlenka neosvědčila. Zkusme však pomyslet alespoň na bezprostřední důsledky: místo optimalizačního algoritmu bychom zřejmě dospěli k algoritmu hledajícímu jen uspokojivá, dostatečná řešení (Varela používá termín *satisficing*: z *satisfy*, uspokojit, splnit, *suffice*, dostačovat); byl by kladen ještě větší důraz na průběžné výsledky řešení a funkce *fitness* by byla prolnuta do celého algoritmu, nevystupovala by v něm jako jednoznačné matematické zobrazení. Je zřejmé, že využití takového enaktivně-evolučního algoritmu by vyžadovalo přeformulování některých problémů, nejednalo by se o přímou náhradu běžných evolučních algoritmů. Museli bychom být připraveni na to, že pro některé přirozeně optimalizační problémy by takový přístup fungoval hůře. Jak často ale potřebujeme skutečně optimální řešení? A jak často se nám ho daří v rozumném čase nalézt?

Tento myšlenkový experiment ukazuje směr, kterým by enaktivní přístup mohl skutečně inspirovat nové metody v počítačové vědě. Neočekávejme, že nám umožní vždy získat lepší řešení přesně těch problémů, které jsme dosud řešili. Na úsvitu oboru umělé inteligence Alan Turing zformuloval svou klíčovou otázku tak, aby vyhovovala jím předpokládaným metodám. Tento směr výzkumu přinesl obrovský rozvoj, byť Turingovým testem za rozumných podmínek žádná umělá inteligence zatím neprošla. Stejně tak enaktivní přístup nebo princip vtělené mysli nám nenabízejí žádné okamžité odpovědi, spíš kladou nové otázky a dávají návod, jak problémy formulovat.

Reference

V citacích jsem používal svůj překlad. Pokud vím, žádná z citovaných anglicky psaných knih přeložena nebyla. Ze zmíněných knih vyšel v češtině Dawkinsův *Selfish Gene* (*Sobecký gen*, přel. V. Kopský, Mladá fronta, 1998), za zmínku stojí také nedávný český překlad knihy J. Haywarda a F. Varely *Gentle Bridges* (*Mosty k porozumění*, přel. M. Šašma, DharmaGaia, 2009), která prezentuje rozhovory vedené v roce 1987 mezi západními vědci, mezi jinými Varelou a Roschovou, a Dalajlámou, a předznamenává tak pozdější knihu *The Embodied Mind*.

V následujícím seznamu uvádím pro přehlednost první vydání (u [2] a [8] vedle aktuálního upraveného vydání, které cituji).

[1] Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[2] Richard Dawkins. *The Extended Phenotype: The Long Reach of the Gene*. Oxford University Press, 1999 – upravené vydání, první vydání 1982. ISBN 978-0-19-288051-2.

[3] Daniel C. Dennett. Review of F. Varela, E. Thompson and E. Rosch, *The Embodied Mind*. *American Journal of Psychology*, 106:121–6, 1993.

[4] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, 1989. ISBN 0201157675.

[5] Ivan M. Havel. Zjednaný svět. *Vesmír*, 78:363, 1999. URL `http://www.vesmir.cz/clanek/zjednany-svet`.

[6] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[7] Richard C. Lewontin, Steven Rose, a Leon J. Kamin. *Not in Our Genes: Biology, Ideology, and Human Nature*. Pantheon Books, New York, 1984. ISBN 0394728882.

[8] Donald A. Norman. *The Design of Everyday Things*. Basic Books, 2002 – aktualizované vydání, první vydání 1988 jako *The Psychology of Everyday Things*. ISBN 0465067107.

[9] Michael O'Neill a Conor Ryan. *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Springer, 2003.

[10] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950. URL `http://www.jstor.org/pss/2251299`.

[11] Francisco J. Varela, Evan Thompson, a Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press, 1991. ISBN 0262220423.

[12] Wikipedia. Cognitive science — Wikipedia, The Free Encyclopedia, 2010. URL `http://en.wikipedia.org/w/index.php?title=Cognitive_science&oldid=363798404`. [Online; accessed 24-May-2010].