# Kognitivní věda a umělé vědomí

Průvodce lidskou myslí a otázkou umělého vědomí

*Jiří Stibor · Deset kapitol pro veřejnost · připraveno s podporou AI*

*18. října 2026 · redakčně upraveno 4. listopadu 2026*

# Úvod k průvodci

O lidské mysli často mluvíme nejsměleji právě tehdy, když se setkáme s novou technologií. Plynulá odpověď počítače vyvolává otázku porozumění, osobní archiv představu digitálního pokračování a zpráva o nové architektuře příslib vědomého stroje. Tento sborník nabízí způsob, jak se v takových tvrzeních orientovat bez technického vzdělání a bez předem předepsaného filosofického názoru.

Deset kapitol sleduje cestu od vlastní zkušenosti přes myšlení a identitu k umělé inteligenci, teoriím vědomí a odpovědnosti za jejich možné důsledky. Kognitivní věda zde znamená mezioborové zkoumání poznávání: propojuje mimo jiné psychologii, neurovědu, filosofii, jazykovědu a výpočetní modelování. Umělé vědomí označuje hypotetickou možnost subjektivního prožívání v uměle vytvořeném systému. Samotný název nepotvrzuje, že takový systém již máme.

Příspěvky lze číst postupně i jednotlivě. Každý má anotaci, klíčová slova, vlastní argumentaci a seznam literatury. Odborné pojmy jsou vysvětleny při prvním důležitém použití; na konci svazku je společný slovníček s odkazy na příslušné kapitoly. Původní otázky kurzu zůstaly zachovány, ale text je přepracován pro širší veřejnost.

Při čtení rozlišujeme tři roviny. Empirický závěr se opírá o pozorování nebo experiment za určitých podmínek. Filosofická teze vysvětluje pojmy, vztahy nebo předpoklady a musí obstát v argumentaci. Normativní stanovisko říká, jak bychom měli jednat; potřebuje také hodnotové důvody. Všechny tři roviny mohou být dobře zdůvodněné, žádná však automaticky nenahrazuje ostatní.

Modelové situace a malé pokusy nejsou vydávány za nové vědecké studie. Pomáhají odhalit předpoklad nebo porovnat výklady. U některých úloh je smysluplnější jasně formulovaná námitka než rychlá odpověď. Poznámky pod čarou vysvětlují drobnosti a občas odlehčují výklad; jejich nadsázka není odborným důkazem.

Literatura zahrnuje historické argumenty i novější výzkumné práce. Datum publikace je součástí významu zdroje: z výsledku určitého modelu nelze bez dalšího vyvozovat vlastnosti všech pozdějších systémů. Preprint je výzkumný text zveřejněný před standardním časopiseckým posouzením nebo mimo něj; jeho existence není potvrzením všech závěrů. Tento svazek je výkladem vybraných problémů, nikoli průběžným žebříčkem produktů.

Hlavním cílem není rozhodnout, zda máme stroje milovat nebo se jich bát. Je jím schopnost poznat, co se vlastně tvrdí, jaká opora je k dispozici a co bychom potřebovali vědět pro silnější závěr. Taková schopnost zůstává užitečná i tehdy, když se technologie a teorie promění.

# Obsah

**Část I Lidská zkušenost a já**

1 Co je vědomí – a proč je tak zatraceně těžké ho definovat ..... 4

2 Mozek jako stroj na predikci ..... 7

3 Mysl, já a iluze identity ..... 10

**Část II Strojová inteligence a vědomí**

4 Umělá inteligence ≠ umělá mysl ..... 13

5 Může stroj myslet? (Turing, Searle, Dennett) ..... 16

6 Umělé vědomí: teoretické modely ..... 19

**Část III Hranice člověka a odpovědnost**

7 Kybernetická identita a rozšířená mysl ..... 22

8 Etika umělého vědomí ..... 25

9 Posthumanismus a transhumanismus ..... 28

10 Závěrečná syntéza: Máme vůbec chtít umělé vědomí? ..... 31

Společný terminologický slovníček ..... 34

Odkazy v obsahu vedou na příslušné příspěvky. Literatura je uvedena na konci každého z nich.

# Část I Lidská zkušenost a já

Vědomí, vnímání a otázka osobní identity.

## 1 Co je vědomí – a proč je tak zatraceně těžké ho definovat

Co znamená něco prožívat


### Anotace

Příspěvek rozlišuje vědomou zkušenost, bdělost, pozornost a inteligenci. Vychází z běžných situací, v nichž se tyto vlastnosti rozcházejí, a vysvětluje, proč měření chování samo neřeší otázku prožitku. Představuje filosofický těžký problém vědomí a možnosti nepřímého vědeckého zkoumání. Jeho závěrem je, že nejistota o povaze zkušenosti není důvodem opustit výzkum ani oprávněním připisovat vědomí všemu, co přesvědčivě odpovídá.



### Klíčová slova

vědomí; subjektivní zkušenost; bdělost; pozornost; inteligence; těžký problém


### Ranní probuzení a noční život

Probudíme se z nepříjemného snu a chvíli ještě hledáme nebezpečí, které v ložnici není. Před okamžikem jsme spali, a přesto jsme něco prožívali. O hodinu později sedíme v tramvaji s otevřenýma očima a nezaznamenáme známého, který nám mává. A večer nám počítač napíše, že rozumí naší únavě. Ve všech třech situacích se nabízí slovo vědomí. Pokaždé však potřebujeme položit trochu jinou otázku.

Bdělost označuje stav organismu, jeho připravenost reagovat. Pozornost znamená výběr nebo zvýhodnění některých informací. Inteligenci obvykle spojujeme se schopností učit se, řešit úlohy a přizpůsobovat se. Subjektivní zkušenost naproti tomu znamená, že určitý stav má pro někoho prožívanou podobu: něco bolí, něco chutná, něco připomíná domov. Tyto vlastnosti spolu souvisejí, ale není možné je bez vysvětlení zaměňovat.

Výzkum snění ukazuje, proč je takové rozlišení užitečné. Siclariová a spoluautoři (2017) porovnávali zprávy po probuzení se záznamy mozkové aktivity. Zkušenosti se objevovaly v různých fázích spánku. Výpověď o snu a elektrický signál poskytovaly odlišné, vzájemně se doplňující informace. Pozorování mozku tak pomáhá zkušenost zkoumat, aniž by se samo proměnilo ve čtení snu z pohledu snícího.

### Co znamená mít zkušenost

Když oba řekneme, že vidíme červenou, můžeme porovnat, jakou barvu ukazujeme, jak ji pojmenujeme a co dělá naše zraková soustava. Přesto si nevyměníme vlastní pohled na svět stejně snadno jako fotografii. Fenomenální vědomí je odborné označení právě pro tuto prožívanou stránku. Slovo fenomenální zde neznamená mimořádné nebo úžasné; patří sem i docela obyčejná nuda.[1]

Block (1995) odlišil tuto rovinu od přístupového vědomí, tedy dostupnosti obsahu pro usuzování, řeč a řízení jednání. Rozdíl lze přiblížit otázkami: co člověk zažil a s čím dokáže dále pracovat?

1 Odborný slovník dokáže dodat slavnostní název i čekání na opožděný autobus. Přejmenování však zkušenost nevysvětluje.

Nemusíme předpokládat dva oddělené orgány. Potřebujeme jen nepovažovat schopnost odpovědět za úplnou definici toho, co má být vysvětleno.

Výpovědi druhých přitom nejsou bezcenné. Když někdo řekne, že ho bolí zub, obvykle to bereme vážně. Jeho slova hodnotíme v souvislosti s chováním, okolnostmi a znalostí lidského těla. Nepřímý důkaz může být velmi přesvědčivý. To, že cizí zkušenost nemáme bezprostředně k dispozici, neznamená, že o ní nemůžeme vědět nic.

### Proč správná odpověď ještě není celé vysvětlení

Chalmers (1995b) nazval těžkým problémem otázku, proč jsou fyzické procesy doprovázeny zkušeností. Vysvětlit rozlišení barev, zapamatování informace nebo tvorbu odpovědi je nesmírně náročné. Přesto podle jeho argumentu zůstává další otázka: proč se něco z toho také nějak prožívá? Toto vymezení je filosofickým stanoviskem o povaze vysvětlení, nikoli laboratorním nálezem nehmotné duše.

Námitka zní, že jsme problém vytvořili způsobem, jakým o sobě přemýšlíme. Frankish (2016) rozvíjí iluzionismus: možná nám introspekce, tedy pozorování vlastních stavů, představuje zkušenost jako něco, čím ve skutečnosti není. Stanovisko je sporné a neznamená jednoduše, že lidé nevnímají nebo že lze ignorovat jejich bolest. Žádá jiné vysvětlení toho, proč se nám některé vlastnosti zkušenosti zdají samozřejmé.

Ani slovem iluze však nelze spor ukončit. Pokud se člověku něco jeví, zůstává třeba vysvětlit vznik tohoto jevení, jeho podobu a souvislost s jednáním. Podobně věta „všechno dělá mozek“ určuje oblast hledání, ale ještě nevysvětluje vztah mezi konkrétní činností mozku a tím, jaké je být daným člověkem. Opatrnost je potřebná na obou stranách sporu.

### Jak pracuje věda s něčím subjektivním

Vědecký výzkum spojuje výpovědi, chování, mozková měření a někdy také řízené zásahy. Hledá pravidelnosti a porovnává soupeřící vysvětlení. Seth a Bayne (2022) ukazují, že teorie vědomí sledují různé cíle: některé především dostupnost obsahu, jiné celkový stav nebo vlastnosti zkušenosti. Rozdílná otázka může vyžadovat rozdílné měření.

Představme si člověka, který nedokáže stisknout tlačítko. Z chybějícího stisku bez dalšího nepoznáme, zda neviděl podnět, nerozuměl pokynu, nebo nemohl pohnout rukou. Výzkumník musí tyto možnosti rozlišit. Stejná úloha u člověka a u stroje navíc nemusí poskytovat stejně silný důkaz: správná věta může vznikat různými mechanismy.

### Malé pozorování vlastní pozornosti

Rozhlédněte se po místnosti a třicet sekund hledejte předměty jedné barvy. Potom se od místnosti odvraťte a zkuste vyjmenovat předměty jiné barvy. Zapište, čím jste si jisti a co pouze odhadujete. Cvičení může ukázat omezení výběru a vybavení, nikoli přesně určit, co všechno jste vědomě viděli.

Při výkladu výsledku uvažujte alespoň dvě možnosti. Některé věci jste mohli přehlédnout; jiné jste zaznamenali, ale už si je nevybavujete. Domácí pozorování tyto možnosti samo nerozhodne. Jeho smyslem je naučit se nepovažovat vlastní jistotu za hotovou teorii mysli.

### Otázky k promyšlení

Je vědomí vlastnost, proces, nebo iluze?

Lze vědomí vůbec objektivně zkoumat?

První otázka nenabízí tři nutně neslučitelné možnosti: vlastnost systému může vznikat probíhajícím procesem a některý její sebepopis může být zavádějící. U druhé otázky pomůže oddělit veřejně kontrolovatelný postup od požadavku, aby badatel přímo sdílel cizí zkušenost.

## Závěr

Začít rozlišením pojmů znamená zabránit tomu, aby jednu otázku nenápadně zodpověděla jiná. Úspěšná komunikace, otevřené oči ani složitá činnost samy nepředstavují úplné vysvětlení prožitku. Současně máme dobré důvody zkušenost zkoumat nepřímo. Právě mezi těmito dvěma body se pohybuje celý sborník.

## Literatura


BLOCK, N. 1995. On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227–247. https://www.nedblock.us/papers/1995_Function.pdf

CHALMERS, D. J. 1995b. Facing Up to the Problem of Consciousness. Journal of Consciousness Studies, 2(3), 200–219. https://consc.net/papers/facing.html

FRANKISH, K. 2016. Illusionism as a theory of consciousness. Journal of Consciousness Studies, 23(11–12), 11–39. https://www.keithfrankish.com/illusionism/

SETH, A. K., A BAYNE, T. 2022. Theories of consciousness. Nature Reviews Neuroscience, 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4

SICLARI, F., BAIRD, B., PEROGAMVROS, L., ET AL. 2017. The neural correlates of dreaming. Nature Neuroscience, 20, 872–878. https://doi.org/10.1038/nn.4545

# 2 Mozek jako stroj na predikci

Jak očekávání vstupuje do vnímání


### Anotace

Studie představuje prediktivní zpracování jako rodinu přístupů, podle nichž se vnímání opírá o průběžné odhadování příčin smyslových podnětů. Vysvětluje rozdíl mezi očekáváním, chybou předpovědi a učením. Na příkladech ukazuje význam nejistoty, těla a aktivního získávání informací. Teoretický rámec odlišuje od jeho konkrétních experimentálních testů a od mnohem silnějšího tvrzení, že samotná predikce vytváří vědomí.



### Klíčová slova

prediktivní zpracování; očekávání; vnímání; nejistota; učení; emulace mozku


### Hrnkem počínaje

Zvednete hrnek, o němž předpokládáte, že je plný. Ukáže se, že je prázdný, a ruka vyletí výš, než jste chtěli. Drobná příhoda připomíná, že jednání nezačínáme pokaždé od nuly. Předem odhadujeme hmotnost věcí, pokračování věty i to, odkud přijde známý zvuk. Teprve nesoulad nás upozorní na některé předpoklady, které běžně zůstávají v pozadí.

Prediktivní přístupy tuto myšlenku rozvíjejí do modelů. Mozek podle nich nečeká pasivně na hotový obraz světa. Průběžně vytváří odhady příčin přicházejících podnětů a upravuje je podle nesouladu s tím, co získává ze smyslů. Clark (2013) předkládá široký výklad tohoto rámce; jednotlivé modely se ovšem liší v tom, jaké mechanismy odhadování předpokládají.

Slovo predikce tu nemusí znamenat vědomé věštění budoucnosti. Může označovat odhad toho, co právě způsobuje určitý zvuk nebo obraz. Když ve špatném světle rozpoznáváme kabát na věšáku, mozek podle tohoto pojetí odhaduje současnou příčinu podnětu. Nemusí k tomu uvnitř existovat další člověk, který si říká svou hypotézu nahlas.

### Kterému údaji věřit

Očekávání má různou váhu. Pokud obvykle necháváme klíče na jedné poličce, hledáme nejprve tam. Jasný pohled na prázdnou poličku však tuto domněnku rychle oslabí. V pološeru si můžeme být méně jisti. Podstatný tedy není jen rozdíl mezi očekáváním a údajem, ale také jejich spolehlivost.

Ve statistických modelech tento vztah zachycuje práce s nejistotou. Předchozí odhad se kombinuje s novým pozorováním; spolehlivějšímu zdroji přísluší větší váha. Ernst a Banks (2002) v konkrétních úlohách ukázali, jak lidé při spojování zrakových a hmatových informací zohledňují jejich relativní spolehlivost. Výsledek podporuje určitý princip kombinování informací, nikoli tvrzení, že každý lidský úsudek probíhá optimálně.

Není také nutné předpokládat, že mozek doslova vypisuje rovnice. Matematický model může zachytit vztah mezi podmínkami a odpovědí, i když biologický mechanismus pracuje jinak než výpočet na papíře. Rozlišení popisu a mechanismu chrání teorii před tím, aby se z užitečného přirovnání stala údajně úplná fotografie mozku.

### Co se děje při překvapení

Chyba předpovědi je nesoulad mezi odhadovaným a získaným vstupem. Není to automaticky omyl člověka ani jeho pocit překvapení. Odchylku lze zpracovat, aniž si ji člověk uvědomí. Naopak

vědomé překvapení může zahrnovat mnoho dalších procesů, včetně významu události pro naše plány.

Představme si, že přítel neodpoví na obvyklý pozdrav. Můžeme opravit odhad situace: neslyšel nás. Můžeme změnit jednání: pozdravíme hlasitěji. Nebo po opakovaných zkušenostech upravíme dlouhodobější očekávání. Tyto kroky odpovídají různým úlohám, které je třeba odlišovat: odhadu právě probíhající situace, získávání dalších údajů a učení do budoucna.

Zjednodušené heslo, že mozek chce jen odstranit překvapení, svádí k námitce: proč tedy nezůstaneme v tiché tmavé místnosti? Organismus však má také potřeby a očekávání týkající se vlastního těla a jednání. Rámce aktivní inference zahrnují právě vztah odhadování a činnosti (Friston, 2010). Tma nevysvětluje, jak se najíme nebo jak získáme informaci, kterou potřebujeme.

### Tělo není pouhá schránka

Vnímání závisí na tom, co můžeme udělat. Nejasný předmět obejdeme, přiblížíme se k němu nebo jej opatrně vezmeme do ruky. Tím měníme další vstup. Wilsonová (2002) rozlišuje více významů tělesně zakotveného poznávání; není to jedna jednoduchá teze, podle níž by libovolný pohyb vysvětlil všechny vlastnosti mysli.

Pro běžného čtenáře je důležité, že tělo poskytuje soustavnou zpětnou vazbu. Únava mění možnosti jednání, poloha hlavy mění obraz a potřeba odpočinku mění význam prostředí. Přidat stroji kameru proto ještě neznamená vytvořit lidský vztah ke světu. Je třeba popsat, jak vstupy, potřeby a jednání skutečně souvisejí.[2]

### Pokus s neúplnou větou

Doplňte zvlášť dvě věty: „Na zahradě štěkal…“ a „Na hladině plaval…“. Potom si představte, že poslední slovo slyšíte velmi nezřetelně. Jaké pokračování byste čekali? Změňte pouze začátek věty a sledujte, které možnosti vás nyní napadají.

Vliv kontextu dokážou vysvětlovat různé modely. Přesnější experiment by potřeboval kontrolovat slyšitelnost, předchozí zkušenost i způsob odpovědi. Dobrá teorie musí předpovědět něco konkrétnějšího než to, že minulost nějak ovlivňuje přítomnost.

### Kde končí užitečnost metafory

Předpovídání je důležitá funkce, ale vztah k prožitku zůstává další otázkou. Aplikace může odhadovat počasí, aniž tím prokazuje, že se těší na slunce. Také u mozku potřebujeme vysvětlit, proč by určitá organizace předpovědí měla být doprovázena zkušeností. Samotné společné slovo predikce nespojuje člověka a jazykový model ve všech ostatních ohledech.

Podobně technický plán emulace celého mozku rozlišuje získání dat, jejich zpracování a modelování relevantních dějů (Sandberg a Bostrom, 2008). Filosofický argument pro zachování zkušenosti při zachování vhodné organizace předkládá Chalmers (1995a). Plán a argument nejsou zprávou o uskutečněném digitálním člověku.

### Otázky k promyšlení

Kdybychom nasimulovali mozek do posledního neuronu, vznikne vědomí?

Je vědomí výpočet?

2 Kamera na kolečkách není hotovým člověkem stejně jako recept v kapse není obědem. Přirovnání může naznačit potřebné části, nenahradí však vysvětlení jejich spolupráce.

U obou otázek napište, jakou shodu mezi původním systémem a modelem požadujete. Shoda odpovědí, dynamiky a fyzických příčin jsou různé nároky. Pak určete, který další předpoklad vás vede od této shody k tvrzení o zkušenosti.

## Závěr

Prediktivní přístupy pomáhají porozumět tomu, proč vnímání potřebuje minulou zkušenost, nejistotu i jednání. Jejich hodnota roste s přesností předpovědí a ochotou připustit nepříznivý výsledek. Označení mozku za predikční stroj je začátkem výzkumné otázky; samo ještě nevyřešilo otázku vědomí.

## Literatura


CHALMERS, D. J. 1995a. Absent Qualia, Fading Qualia, Dancing Qualia. In T. Metzinger (ed.), Conscious Experience (s. 309–328). Imprint Academic. https://consc.net/papers/qualia.html

CLARK, A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181–204. https://doi.org/10.1017/S0140525X12000477

ERNST, M. O., A BANKS, M. S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415, 429–433. https://doi.org/10.1038/415429a

FRISTON, K. 2010. The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11, 127–138. https://doi.org/10.1038/nrn2787

SANDBERG, A., A BOSTROM, N. 2008. Whole Brain Emulation: A Roadmap. Technical Report 2008-3. Future of Humanity Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:a6880196-34c7-47a0-80f1-74d32ab98788

WILSON, M. 2002. Six views of embodied cognition. Psychonomic Bulletin & Review, 9, 625–636. https://doi.org/10.3758/BF03196322

# 3 Mysl, já a iluze identity

Jak se měníme a přesto o sobě mluvíme jako o téže osobě


### Anotace

Příspěvek rozlišuje bezprostřední sebeprožívání, životní příběh a otázku přetrvávání osoby. Ukazuje, proč proměna vzpomínek ani nepřesnost sebepopisu samy nedokazují, že člověk neexistuje. Pomocí případu starého deníku a hypotetických kopií zkoumá rozdíl mezi podobností a totožností. Závěr nabízí kritéria pro přesnější rozhovor o lidské i digitální identitě, aniž rozhoduje filosofický spor jedinou metaforou.



### Klíčová slova

minimální já; narativní já; paměť; osobní identita; psychologická kontinuita; sebepojetí


### Cizinec ve vlastním deníku

Otevřeme deník z dospívání a najdeme v něm názory, za které bychom se dnes styděli. Některé události si pamatujeme jinak, jiné vůbec. Přesto běžně řekneme, že deník patří nám. Toto jednoduché přivlastnění obsahuje několik různých vztahů: rozpoznání rukopisu, vzpomínku na minulost, pokračování těla a přijetí životního příběhu.

Kdyby identita znamenala naprostou neměnnost vlastností, nevydržela by ani běžný den. Učíme se, unavujeme se a měníme názory. Otázkou tedy není, zda se měníme, ale které změny jsou slučitelné s pokračováním téže osoby. Filosofie zde hledá kritérium, podle něhož vůbec poznáme, o jaký druh pokračování jde.

Je užitečné odlišit kvalitativní shodu od numerické identity. Dvě knihy mohou mít stejný text a vazbu, přesto jde o dva výtisky. Numerická identita znamená jednu a tutéž věc nebo osobu. Kvalitativní podobnost připouští více různých nositelů podobných vlastností. V diskusi o digitálních kopiích se tento rozdíl stane rozhodujícím.

### Dvě podoby já

Gallagher (2000) rozlišuje minimální a narativní já. Minimální já označuje bezprostřední sebeprožívání: bolest se děje mně, pohyb zakouším jako svůj. Nemusím si přitom vyprávět celý životopis. Narativní já souvisí s příběhem, kterým propojuji minulost, současné vztahy a budoucí záměry.

Narativní já není jen soukromý text v hlavě. K příběhu přispívají druzí lidé, fotografie, rodinné vzpomínky a společenské role. Někdo nás zná jako rodiče, jiný jako kolegu. Rozdílné pohledy se mohou doplňovat, ale také střetávat. Z této mnohosti bez dalšího neplyne, že uvnitř žije několik samostatných osob.

Také minimální já a životní příběh se mohou rozcházet. Člověk nemusí právě přemýšlet o své minulosti, aby cítil teplo slunce. Digitální profil naproti tomu může obsahovat podrobný životopis, aniž tím doloží aktuální sebeprožívání. Databáze vět o mně není automaticky někým, kdo je mnou.

### Paměť není nedotčený archiv

Vzpomínání nám umožňuje navazovat na minulé zkušenosti. To však neznamená, že každé vybavení představuje přehrání nezměněného záznamu. Výzkum popisuje paměť také jako konstruktivní proces, v němž se uplatňuje současný kontext a spojování dostupných prvků

(Schacter a Addis, 2007). Konstruktivní zde neznamená svévolně vymyšlený; znamená, že podoba vzpomínky vzniká činností systému.

Představme si sourozence, kteří se neshodnou, jak proběhla rodinná dovolená. Jeden si vybaví dlouhé čekání, druhý zábavný rozhovor. Neshoda může mít různé příčiny: věnovali pozornost jiným věcem, některé podrobnosti zapomněli nebo je později doplnili. Bez dalších podkladů nelze každému sporu přidělit jediný mechanismus.

Z nepřesnosti paměti navíc nevyplývá, že všechna svědectví mají stejnou cenu. Fotografie, dobové zápisy a nezávislé výpovědi mohou pomoci jednotlivé nároky posuzovat. Stejně tak chybná vzpomínka neodnímá člověku jeho hodnotu ani automaticky neukončuje jeho existenci. Kritérium osobní identity nemáme směšovat s hodnocením spolehlivosti určité výpovědi.

### Kdo zůstává po změně

Biologické pojetí sleduje pokračování organismu. Psychologická pojetí dávají větší význam návaznosti vzpomínek, záměrů a dalších duševních stavů. V běžném životě se tyto vztahy často překrývají, a proto si jejich rozdíl neuvědomujeme. Myšlenkové experimenty je záměrně oddělují, aby ukázaly, co naše odpověď skutečně předpokládá.

Parfit (1971) zkoumá případy větvení psychologické kontinuity. Kdyby dva různí pokračovatelé navazovali stejným způsobem na jednoho předchůdce, nestačí jen uvést společné vzpomínky a prohlásit oba za numericky tutéž osobu. Filosof může dojít k závěru, že pro to, na čem nám v přežití záleží, mají některé vztahy větší význam než samotná identita.

Příkladem z jiné oblasti je Théseova loď, na níž se postupně vyměňují části. Podobnost s člověkem může pomoci otázku otevřít, ale loď nemá lidskou paměť ani prožitek. Ani postupná výměna materiálu proto sama nedokazuje, že osobní identita je vyřešena. Každá analogie přenáší jen některé vztahy.[3]

### Co znamená iluze já

Sebepopis může vytvářet dojem větší jednoty a průhlednosti, než jakou skutečné rozhodování má. To je podnět ke zkoumání mechanismů. Výrok „já je iluze“ však potřebuje upřesnění: má být iluzorní dokonalá kontrola, neměnné vnitřní centrum, nebo samotná existence subjektivní perspektivy? Tyto teze nejsou zaměnitelné.

Námitka proti zjednodušenému odmítnutí já je jednoduchá. To, že člověk svůj život vykládá pomocí modelu, ještě nedokazuje neexistenci toho, co modeluje. Mapa města může být neúplná a město přesto existuje. Také tato analogie má mez: neříká, jaký druh skutečnosti já je. Brání pouze neplatnému přechodu od nepřesného sebepopisu k úplnému zániku osoby.

### Práce s vlastním příběhem

Vyberte neškodnou událost, kterou zachycuje starý zápis nebo fotografie. Nejdříve ji krátce popište zpaměti, teprve potom se podívejte na záznam. Oddělte shodu, rozdíl a to, co záznam vůbec nemůže rozhodnout. Fotografie může ukázat přítomnost člověka, ale ne všechny jeho tehdejší pocity.

Smyslem není určit, zda máte dobrou nebo špatnou paměť. Sledujte, které závěry skutečně podporuje porovnání a které jste přidali sami. Cvičení můžete vynechat nebo použít smyšlený příběh; pro porozumění pojmům není třeba otevírat bolestné osobní vzpomínky.

---

3 Při rodinném setkání zpravidla neověřujeme totožnost hostů soupisem původních atomů. Přesto nám záleží na tom, kdo přišel. Běžná praxe je bohatší než jednoduché heslo o neměnnosti materiálu.

## Otázky k promyšlení

Když se změní vzpomínky, jsem to pořád já?

Co přesně by musela AI mít, aby měla „identitu“?

U první otázky určete, které kritérium pokračování používáte. U druhé rozlište stálé označení systému, návaznost jeho chování, model sebe sama a subjektivní perspektivu. Jedna vlastnost nemusí dokazovat ostatní.

## Závěr

Já není jeden snadno přemístitelný údaj. Sebeprožívání, životní vyprávění a přetrvávání osoby se navzájem ovlivňují, ale vyžadují rozdílné otázky. Přesnější slovník nám dovoluje uznat skutečné proměny člověka, aniž z každé změny vytváříme novou osobu nebo každou podobnost vydáváme za pokračování původního života.

## Literatura

GALLAGHER, S. 2000. Philosophical conceptions of the self: Implications for cognitive science. Trends in Cognitive Sciences, 4(1), 14–21. https://doi.org/10.1016/S1364-6613(99)01417-5

PARFIT, D. 1971. Personal Identity. The Philosophical Review, 80(1), 3–27. https://doi.org/10.2307/2184309

SCHACTER, D. L., A ADDIS, D. R. 2007. The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. Philosophical Transactions of the Royal Society B, 362(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

# Část II Strojová inteligence a vědomí

Schopnosti umělých systémů, filosofické argumenty a vědecké modely.

## 4 Umělá inteligence ≠ umělá mysl

Proč nás plynulý rozhovor tak snadno přesvědčí


### Anotace

Příspěvek rozlišuje výkon umělé inteligence, porozumění a subjektivní zkušenost. Přístupně vysvětluje vznik odpovědi jazykového modelu a význam zadání, naučených parametrů a dodaných podkladů. Na modelové situaci z veřejné knihovny ukazuje, proč přesvědčivý jazyk vyžaduje ověřování. Nevyvozuje neexistenci mysli z umělého původu; požaduje, aby každému tvrzení odpovídal vhodný důkaz.



### Klíčová slova

umělá inteligence; jazykový model; porozumění; význam; důvěra; antropomorfizace


### Hlas za přepážkou

Představme si, že městská knihovna nabízí textového asistenta. Zeptáme se na knihu, asistent odpoví vlídně, doporučí související četbu a dodá: „Mám radost, že vás toto téma zajímá.“ Naše zkušenost s lidmi nás vede k tomu, abychom podobnou větu četli jako výraz vnitřního stavu. U stroje však potřebujeme zjistit, jak odpověď vznikla a jaký závěr z ní skutečně plyne.

Není sporné, že rozhovor může být užitečný. Systém může najít relevantní titul nebo pomoci formulovat otázku. Výkon v této úloze ovšem není totožný s prožitkem radosti ani s celkovým porozuměním situaci čtenáře. Znak nerovnosti v názvu kapitoly má připomenout právě toto rozlišení. Není důkazem, že umělá mysl je nemožná.

### Jak vzniká další slovo

Velký jazykový model se během trénování učí vztahy v rozsáhlých datech. U běžných autoregresivních modelů pak při generování vzniká text postupně podle dostupného kontextu. Základní jednotkou přitom nemusí být celé slovo, ale token, například jeho část. Naučené číselné parametry ovlivňují, které pokračování bude model upřednostňovat.

Tento stručný popis neznamená, že model pouze vytahuje hotovou větu z kartotéky. Naučené vztahy mu mohou umožnit kombinovat prvky novým způsobem a zvládat náročné úlohy. Zároveň mechanismus generování sám neposkytuje záruku pravdivosti. Pravděpodobně působící pokračování může obsahovat smyšlenou publikaci i velmi přesné vysvětlení; konkrétní výsledek potřebuje kontrolu.

Architektura transformeru, kterou představili Vaswani a spoluautoři (2017), využívá mechanismus attention. Ten umožňuje určitým způsobem vážit vztahy mezi reprezentacemi. Název překládáme jako pozornost, ale tím jsme ještě neprokázali vědomé soustředění. Technické názvy často přebírají lidská slova, aniž přebírají celý jejich psychologický význam.[4]

4 Také počítač může mít „paměť“, aniž si při prohlížení starých souborů povzdechne nad mládím. Společné slovo usnadňuje popis funkce, ale umí do něj přidat nezasloužený příběh.

### Co znamená dodat modelu podklady

Vraťme se do knihovny. Asistent může odpovídat pouze z toho, co se naučil při trénování, nebo může mít k dispozici aktuální katalog. Tyto situace nejsou stejné. Pokud katalog obsahuje informaci o dostupnosti knihy, systém ji může použít při odpovědi. Není však rozumné předpokládat, že každá dobře formulovaná věta katalog skutečně věrně reprodukuje.

Rozdíl mezi naučenými parametry a právě dodaným kontextem připomíná rozdíl mezi dlouhodobou průpravou a podklady položenými na stůl. Analogie není přesným modelem lidské paměti. Pomáhá jen pochopit, proč doplnění dokumentu nemusí znamenat přetrénování systému. Brown a spoluautoři (2020) zkoumali využívání příkladů v kontextu na konkrétním historickém modelu; výsledek nepopisuje automaticky všechny další aplikace.

### Rozumět může znamenat více věcí

Porozuměním někdy myslíme schopnost správně používat vztahy: odpovědět na otázku, opravit omyl nebo uplatnit pravidlo v nové situaci. Jindy míníme vztah slov ke světu. A někdy máme na mysli prožívaný okamžik pochopení. Spor může vzniknout už tím, že každý účastník používá jiný význam.

Benderová a Koller (2020) upozorňují na rozdíl mezi jazykovou formou a významem. Jejich argument vyzývá k opatrnosti při odvozování porozumění z práce s textovými pravidelnostmi. Nelze jej však bez rozboru podmínek zaměnit za empirický důkaz, že žádný myslitelný umělý systém nikdy nemůže rozumět. Systém zapojený do vnímání, jednání a komunikace představuje další otázky.

Také opačné tvrzení potřebuje oporu. To, že model zvládne úlohu, neukazuje samo, co jeho výkon umožnilo. Mohl využít přenositelné pravidlo, podobný známý příklad nebo nápovědu ve formulaci. Rozlišit tyto možnosti je možné pomocí vhodných změn zadání. Z jednotlivého úspěchu ani selhání však nelze jednoduše vyčíst úplnou povahu jeho mysli.

### Důvěra potřebuje konkrétní předmět

Představme si, že asistent doporučí údajnou knihu včetně autora, roku a citátu. Působí to ověřitelněji než neurčité doporučení. Pokud je však bibliografický záznam smyšlený, podrobnost pouze zesílila dojem jistoty. Výzkumné hodnocení TruthfulQA sleduje, jak mohou modely vytvářet nepravdivé odpovědi v určitých typech otázek (Lin, Hilton a Evans, 2022). Plynulost a pravdivost jsou různé vlastnosti.

Termín halucinace modelu v tomto kontextu označuje nepodložený nebo nepravdivý výstup. Nepřipisuje stroji smyslovou halucinaci. Pro čtenáře je praktické ptát se, zda lze tvrzení doložit nezávislým zdrojem, zda systém rozlišuje dokument a vlastní odhad a zda připouští, že potřebný údaj chybí.

Silná důvěra navíc nemusí být důvěrou ve všechno. Můžeme systému svěřit návrh názvů pro čtenářský klub a zároveň ověřovat údaje o skutečných knihách. Takové rozlišení není nerozhodnost; odpovídá různým následkům chyby. Užitečnost služby se má hodnotit vzhledem k tomu, k čemu ji používáme.

### Malý pokus s pravidly knihovny

Vytvořte fiktivní pravidlo: člen si může půjčit nanejvýš čtyři knihy a nemůže si půjčit další, pokud má nevrácenou knihu po termínu. Připravte tři případy: člověka se třemi knihami bez prodlení, člověka se čtyřmi knihami a člověka, o jehož prodlení nic nevíte. Zkuste odpovědět sami a případně porovnejte odpověď asistenta.

Potom změňte limit na dvě knihy, zachovejte ostatní pravidla a vytvořte odpovídající nové případy. Hodnoťte zvlášť správnost a přiznání chybějící informace. Úloha zkouší práci s dodaným pravidlem. Ani bezchybný výsledek by neprokázal vědomí; chyba by zase nebyla důkazem neexistence jakéhokoli porozumění.

### Otázky k promyšlení

Kdy nás mate jazyk?

Je porozumění nutné pro vědomí?

U první otázky si všimněte první osoby a výrazů emocí. U druhé určete, co nazýváte porozuměním. Samotný prožitek barvy zjevně nepotřebuje odborný výklad optiky; mnohem širší pojetí porozumění by však vyžadovalo samostatný argument.

### Závěr

Rozhovor s AI je skutečnou interakcí, která může mít skutečné následky pro člověka. O vnitřních vlastnostech systému však nerozhoduje pouze jeho komunikační přesvědčivost. Potřebujeme hodnotit schopnost, původ odpovědi a prožitek odděleně. Teprve potom lze vést spor o umělou mysl tak, aby se neopíral jen o dojem z hlasu za přepážkou.

### Literatura

BENDER, E. M., A KOLLER, A. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of ACL, 5185–5198. https://aclanthology.org/2020.acl-main.463/

BROWN, T. B., ET AL. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33. https://arxiv.org/abs/2005.14165

LIN, S., HILTON, J., A EVANS, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of ACL, 3214–3252. https://aclanthology.org/2022.acl-long.229/

VASWANI, A., ET AL. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, 30. https://arxiv.org/abs/1706.03762

# 5 Může stroj myslet? (Turing, Searle, Dennett)

Tři způsoby, jak položit jednu zdánlivě jednoduchou otázku


### Anotace

Studie porovnává Turingovu imitační hru, Searlův čínský pokoj a Dennettův intencionální postoj. Ukazuje, že jde o různé otázky: podle čeho schopnost posuzovat, co pro ni může být dostačující a jak ji vysvětlovat. Funkcionalismus představuje jako rodinu filosofických přístupů, nikoli jako ztotožnění mysli s libovolným programem. Součástí je krátká debata, v níž má čtenář hledat nejsilnější námitku proti vlastnímu názoru.



### Klíčová slova

Turingův test; čínský pokoj; funkcionalismus; intencionální postoj; argument; prožitek


### Nejdříve určit, co se má rozhodnout

Kdyby nám stroj napsal krásný dopis, mohli bychom ocenit jeho text. Kdyby poradil v neznámé situaci, mohli bychom obdivovat jeho schopnost. Kdyby popsal vlastní osamělost, otevřela by se jiná otázka: prožívá něco, nebo jen vytváří vhodnou odpověď? Jediné sloveso myslet tyto různé nároky snadno zakryje.

Je rovněž rozdíl mezi otázkou, jak mysl poznáváme, a otázkou, co ji tvoří. Z toho, že u druhých vycházíme hlavně z chování, neplyne, že mysl není nic než chování. A z neúplnosti našeho přístupu neplyne, že všechny důkazy mají stejnou hodnotu. Na tomto rozlišení stojí velká část následující debaty.

### Turing a veřejně posouditelný výkon

Turing (1950) navrhuje nahradit neurčitou otázku po myšlení rozborem imitační hry. Jeho původní text vychází z rozlišování muže a ženy v textové komunikaci a zvažuje nahrazení účastníka strojem. Později se Turingovým testem běžně rozumí rozhovor, v němž posuzovatel rozlišuje člověka a počítač. Jednotlivé protokoly však mohou být různě náročné.

Přínos spočívá v tom, že místo soudu podle vzhledu posuzujeme konkrétní interakci. Záleží na délce rozhovoru, otázkách i znalostech posuzovatele. Úspěch je proto vždy třeba vztáhnout k podmínkám. Krátký rozhovor o známém tématu není totéž jako dlouhodobá schopnost zvládat nepředvídatelné situace.

Navíc podobnost člověku nemusí být jediným žádoucím cílem. U poradenské služby můžeme více ocenit přesnost a přiznání nejistoty než lidsky působící zaváhání. Turingova metoda otevřela plodnou cestu k posuzování schopností. Sama však neudělala z každého nerozlišeného rozhovoru přímé měření subjektivního prožitku.

### Searle a člověk, který nezná čínštinu

Searle (1980) předkládá myšlenkový experiment. Člověk v pokoji nezná čínštinu, ale má pravidla, podle nichž manipuluje se znaky. Na vstupu dostává otázky, na výstupu poskytuje odpovědi, které mohou vypadat správně. Člověk uvnitř přesto významu nerozumí. Searle tím zpochybňuje tezi, že samotné provádění formálního programu postačuje k porozumění.

Rozdíl mezi syntaxí a sémantikou je rozdílem mezi pravidly zacházení se znaky a jejich významem. Argument není zprávou o skutečně postaveném pokoji ani prostým odmítnutím každé umělé mysli.

Míří na určitý nárok: že správný program sám stačí. Otázka jiných relevantních vlastností systému zůstává otevřená.

Známá systémová námitka odpovídá, že rozumět by mohl celý uspořádaný systém, i když nerozumí jeho jednotlivá část. Písmeno na papíře rovněž samo nevypráví román. Přirovnání ukazuje, proč vlastnost části nelze automaticky převést na celek. Searle na námitku odpovídá a spor pokračuje na úrovni předpokladů o významu a mysli.

### Co má být zachováno při změně materiálu

Funkcionalismus vymezuje mentální stavy prostřednictvím jejich kauzálních rolí: čím jsou vyvolávány, jak se vztahují k dalším stavům a jak ovlivňují jednání. Nejde jen o shodu jedné výsledné věty. Dva systémy mohou dnes odpovědět stejně, ale po změně okolností reagovat zcela jinak.

Block (1978) zkoumá obtíže funkcionalistických teorií a jejich vztah k neobvyklým možným realizacím mysli. Chalmers (1995a) naproti tomu argumentuje pro zachování zkušenosti při zachování vhodné jemné funkční organizace. Spor zkouší, které vlastnosti považujeme za rozhodující.

Pro veřejnou debatu z toho plyne praktický nárok: když někdo řekne, že počítač dělá totéž co člověk, má upřesnit, co znamená totéž. Podobnost navenek, shoda vybraných schopností a shoda celé relevantní organizace jsou odlišné předpoklady. Výsledek sporu může záviset právě na tom, který z nich jsme přijali.

### Dennett a různé způsoby vysvětlení

Dennett (1981) popisuje intencionální postoj: jednání systému předvídáme tím, že mu přisuzujeme přesvědčení a cíle. Řekneme třeba, že šachový program chce ochránit krále. Takový popis může zachycovat stabilní vztahy a být užitečný. Nemusí přitom tvrdit, že program prožívá strach o svého panovníka.

To neznamená, že všechny řeči o mysli jsou pouhé ozdoby. Máme zkoumat, kdy daný způsob vysvětlení skutečně funguje a co vysvětluje. Výraz chce může označovat různé věci: funkční směřování systému, dlouhodobý projekt osoby nebo pociťovanou touhu. Plynulý přechod mezi nimi není argumentem pro jejich totožnost.

Dennett a Kinsbourne (1992) také kritizují představu vnitřního diváka, kterému mozek promítá hotový svět. Pokud uvnitř dosadíme dalšího človíčka, musíme znovu vysvětlit jeho vnímání. Takzvaný homunkulus původní problém přesouvá. Odmítnout jej však neznamená popřít veškerou koordinaci v mozku.[5]

### Debata se změnou stran

Napište krátkou obhajobu výroku: „Systém, který zachová všechny relevantní funkce porozumění, opravdu rozumí.“ Určete význam porozumění a jeden předpoklad, bez něhož závěr neplatí. Potom napište nejsilnější námitku. Nepoužívejte pouze výrazy „jen stroj“ nebo „úplně jako člověk“.

Nakonec strany vyměňte. Zkuste námitku formulovat tak, aby ji její zastánce považoval za věrnou. Cvičení neuděluje body za správnou filosofickou víru. Zkouší, zda rozpoznáme, co druhá strana skutečně tvrdí a kde se rozcházejí naše důvody.

5 Vnitřního diváka nelze zachránit tím, že jej přejmenujeme na vedoucího oddělení vědomí. Stále bychom museli zjistit, kdo se dívá v jeho hlavě.

### Otázky k promyšlení

Stačí chování, nebo je nutný prožitek?

Jak poznáme, že něco „cítí“?

K první otázce doplňte účel: stačí k posouzení služby, inteligence, nebo fenomenálního vědomí? U druhé rozlišujte důkazy a naprostou jistotu. Nepřímé poznání může být silné, ale jeho přenos na jiný druh systému potřebuje obhajobu.

### Závěr

Turing, Searle a Dennett nám nedávají tři jednoduché hlasy v anketě. Učí nás oddělovat způsob posouzení, podmínky vzniku schopnosti a úroveň vysvětlení. Teprve když víme, o kterou otázku jde, lze smysluplně přejít k výzkumu mechanismů, které by mohly s vědomím souviset.

### Literatura


BLOCK, N. 1978. Troubles with Functionalism. Minnesota Studies in the Philosophy of Science, 9, 261–325. https://hdl.handle.net/11299/185298

CHALMERS, D. J. 1995a. Absent Qualia, Fading Qualia, Dancing Qualia. In T. Metzinger (ed.), Conscious Experience (s. 309–328). Imprint Academic. https://consc.net/papers/qualia.html

DENNETT, D. C. 1981. True Believers: The Intentional Strategy and Why It Works. In A. F. Heath (ed.), Scientific Explanation (s. 150–167). Clarendon Press. https://web.stanford.edu/~paulsko/papers/DennettTB.pdf

DENNETT, D. C., A KINSBOURNE, M. 1992. Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain Sciences, 15(2), 183–201. https://dl.tufts.edu/concern/pdfs/wp988x36n

SEARLE, J. R. 1980. Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

TURING, A. M. 1950. Computing Machinery and Intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

# 6 Umělé vědomí: teoretické modely

## Co by měl umět dobrý test


### Anotace

Příspěvek srovnává globální pracovní prostor, teorii integrované informace a tělesně zakotvené přístupy. Přibližuje jejich rozdílné otázky pomocí příkladů sdílení informací, příčinných vztahů a aktivního vnímání. Vysvětluje význam experimentální předpovědi a meze číselných ukazatelů. Současné modely představuje jako nástroje výzkumu, nikoli jako hotové návody k vytvoření prožívajícího stroje.



### Klíčová slova

globální pracovní prostor; integrovaná informace; tělesné zakotvení; indikátor; měření; emergence


### Když se neshodneme, co hledáme

Jedna výzkumná skupina hledá způsob, jak se informace stává dostupnou mnoha procesům. Druhá zdůrazňuje vnitřní příčinnou organizaci systému. Třetí zkoumá vztah organismu k tělu a prostředí. Všechny mohou mluvit o vědomí, přesto nemusí předpovídat totéž. Přehled Setha a Bayna (2022) upozorňuje právě na rozdílné cíle a předpoklady teorií.

Teorie navrhuje obecné vztahy. Model je zpřesňuje například architekturou nebo rovnicemi. Indikátor je pozorovatelný znak, z něhož o určité vlastnosti usuzujeme. Měřitelná veličina se nestává přímým přístupem ke zkušenosti jen proto, že má číselnou hodnotu. Její význam závisí na zdůvodnění a na podmínkách, v nichž byla ověřena.

### Informace, která je dostupná celku

Baarsova teorie globálního pracovního prostoru, označovaná zkratkou GWT, vychází z rozdílu mezi lokálním zpracováním a širokou dostupností vybraného obsahu (Baars, 2005). Informace se může stát použitelnou pro plánování, zapamatování i odpověď. Představa není totožná s malým divadlem, v němž sedí vnitřní pozorovatel.

Pomůže omezená analogie s organizací oslavy. Jeden člověk ví, že změnili čas, jiný připravuje jídlo a další objednává dopravu. Dokud první informace zůstává soukromá, ostatní činnosti se nepřizpůsobí. Zpřístupnění údaje promění koordinaci. Příklad objasňuje funkční přínos sdílení informací.

Je třeba odlišit obecnou GWT od globálního neuronálního pracovního prostoru, GNW či GNWT. Druhý název označuje neuronální rozpracování s konkrétními mechanismy a předpověďmi (Dehaene, Kerszberg a Changeux, 1998). Pouhé vytvoření společného úložiště v počítači tedy ještě neznamená, že jsme ověřili všechny teze těchto přístupů.

### Vnitřní vztahy nejsou objemem souboru

Teorie integrované informace, IIT, spojuje vědomí s vnitřní kauzální strukturou fyzického systému. Ve verzi IIT 4.0 hrají roli příčiny, účinky a jejich neredukovatelnost při rozdělení systému (Albantakis a spoluautoři, 2023). Je to výrazný teoretický závazek, nikoli všeobecně přijatá definice zkušenosti.

Větší množství dat proto není automaticky větším množstvím vědomí. Dvě zařízení mohou zobrazovat stejně složité obrazce, protože přijímají společné vysílání, aniž se vzájemně ovlivňují.

Jinde se změna jedné části promítá do možností druhé. Rozhodující je zde rozdíl mezi souběhem a vzájemným příčinným působením.

Symbol Φ bývá představován jako číslo vědomí. Bez určení verze teorie, modelu a konkrétní veličiny však takové označení klame. Pro teorii je důležitá fyzická realizace, nejen popis toho, co program napodobuje. Dva systémy se stejným výstupem proto nemusí být z jejího hlediska totožné.

### Proč záleží na možnosti jednat

Tělesně zakotvené a enaktivní přístupy obracejí pozornost k aktivnímu vztahu organismu a prostředí. Nejde o jednu jednotnou teorii. O'Regan a Noë (2001) například zdůrazňují zákonitosti, podle nichž se zrakový vstup mění při jednání, a naše praktické zvládání těchto vztahů.

Když chceme zjistit, co je za hrnkem, změníme polohu hlavy. Ve videu může následovat stejný obraz, ale jeho změnu už neřídíme. Rozdíl může být důležitý pro získání informací a přizpůsobení se nové situaci. Přesto samotný pohyb není důkazem prožitku a nevyřeší zkušenost při snění nebo představování. Teorie musí zahrnout i tyto případy.

### Co přineslo společné porovnání

Studie Cogitate Consortium a spoluautorů (2025) porovnala vybrané předpovědi GNWT a IIT v předem dohodnutém a registrovaném uspořádání. Výzkum s 256 účastníky využil různé způsoby měření mozkové aktivity při sledování viditelných podnětů. Některá očekávání podpořil, jiná zpochybnil; výsledek nebyl jednoznačným vítězstvím jedné celé teorie.

Konkrétně v zadních oblastech mozkové kůry přetrvávala informace rozlišující tváře a předměty po dobu jejich zobrazení. To odpovídalo části předpovědí IIT o trvání vědomého obsahu. Nepotvrdila se však předpovězená trvalá synchronizace mezi zadními oblastmi, tedy časově sladěná aktivita, která měla tento obsah provázet. Pro GNWT představovala problém zejména chybějící nová vlna aktivace při zmizení podnětu, označovaná jako ignition neboli „zažehnutí". Test tak poskytl dílčí podporu i konkrétní námitky; nešlo o jednoduché vyřazení jedné teorie (Cogitate Consortium a spoluautoři, 2025).

Právě předem určené očekávání dává výsledku váhu. Pokud bychom po každém měření změnili význam tvrzení tak, aby vždy souhlasilo, přestalo by rozlišovat mezi možnostmi. Oprava teorie je legitimní, musí však být rozpoznatelná jako oprava. Přenos výsledků z člověka na umělý systém vyžaduje další ověření.

### Lákavost jediného čísla

Casali a spoluautoři (2013) představili index PCI, který zachycuje složitost odpovědi mozku na definovaný zásah. Ve zkoumaných lidských stavech poskytoval užitečné rozlišení bez požadavku na aktuální behaviorální odpověď. Postup zahrnuje specifickou stimulaci a měření; není totožný s výpočtem Φ ani s kompresí libovolného textu.

Přenést podobný výpočet na počítačové výstupy znamená vytvořit novou metriku, jejíž význam je třeba ověřit znovu. Stabilně opakované číslo může být spolehlivé, a přesto měřit jinou vlastnost, než kterou mu připisujeme. Hodiny mohou velmi přesně ukazovat čas a vůbec neměřit naši trpělivost.[6]

6 Desetinná místa přidávají přesnost zápisu, nikoli automaticky přesnost poznání. Zvlášť působivé číslo si zaslouží zvlášť obyčejnou otázku: co jste vlastně měřili?

### Domácí model sdílení

Připravte dvě kartičky. Na první je údaj, zda je otevřená zahrada; na druhé, zda je volný stůl. Návštěvu plánujeme jen tehdy, když platí obě podmínky. Vyzkoušejte všechny čtyři kombinace a potom jednu kartičku zakryjte. Kdy lze odpovědět a kdy je správné přiznat, že informace chybí?

Cvičení ukazuje význam dostupnosti údajů. Pokud víme, že zahrada je zavřená, lze návštěvu odmítnout i bez druhého údaje. Pokud je otevřená a stav stolu neznáme, kladné rozhodnutí není podloženo. Příklad procvičuje úsudek o dostupných údajích.

### Otázky k promyšlení

Je vědomí emergentní vlastnost?

Lze ho měřit?

Emergence může označovat vznik systémové vlastnosti z organizace částí, nebo silnější filosofickou neredukovatelnost. Určete význam dříve, než slovo použijete jako vysvětlení. U měření se ptejte na postup, ověření a rozsah platnosti, nikoli pouze na dostupnost čísla.

### Závěr

Teorie vědomí mají cenu právě tím, že odlišují možné mechanismy a dovolují formulovat předpovědi. Jejich spory nejsou důvodem zaměnit vědu za libovolnou domněnku. Jsou důvodem požadovat přesnější vazbu mezi teorií, výsledkem a tvrzením o umělém systému. Název architektury ani jediná metrika tuto vazbu nenahradí.

### Literatura


ALBANTAKIS, L., ET AL. 2023. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLOS Computational Biology, 19(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

BAARS, B. J. 2005. Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. Progress in Brain Research, 150, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9

CASALI, A. G., ET AL. 2013. A theoretically based index of consciousness independent of sensory processing and behavior. Science Translational Medicine, 5(198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294

COGITATE CONSORTIUM, FERRANTE, O., GORSKA-KLIMOWSKA, U., ET AL. 2025. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature, 642, 133–142. https://doi.org/10.1038/s41586-025-08888-1

DEHAENE, S., KERSZBERG, M., A CHANGEUX, J.-P. 1998. A neuronal model of a global workspace in effortful cognitive tasks. PNAS, 95(24), 14529–14534. https://doi.org/10.1073/pnas.95.24.14529

O'REGAN, J. K., A NOË, A. 2001. A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences, 24(5), 939–973. https://doi.org/10.1017/S0140525X01000115

SETH, A. K., A BAYNE, T. 2022. Theories of consciousness. Nature Reviews Neuroscience, 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4

# Část III Hranice člověka a odpovědnost

Technologické opory, etika a možné budoucnosti lidského života.

## 7 Kybernetická identita a rozšířená mysl

Kam sahá naše poznávání, když používáme druhé lidi a věci


### Anotace

Studie zkoumá rozdíl mezi pomůckou, kognitivním odlehčením a filosofickou tezí rozšířené mysli. Vychází z používání kalendáře, zápisníku a navigace a sleduje, jak tyto prostředky mění naše možnosti i závislosti. Rozlišuje poznávací funkci, digitální reprezentaci a subjektivní zkušenost. Závěr ukazuje, proč lze důležitou pomůcku chránit, aniž jí přisuzujeme vlastní vědomí nebo ji jednoduše ztotožňujeme s člověkem.



### Klíčová slova

rozšířená mysl; kognitivní odlehčení; digitální identita; autonomie; paměť; kyborg


### Den, který drží pohromadě kalendář

Ráno nás telefon upozorní na schůzku, kterou bychom si jinak nevybavili. V obchodě otevřeme seznam a odpoledne podle mapy najdeme neznámou ulici. Nejde jen o pohodlí. Bez těchto prostředků bychom některé činnosti uskutečnili jinak, jiné později a na některé zapomněli. V jakém smyslu jsou takové věci součástí našeho poznávání?

Jedna odpověď říká, že mysl zůstává uvnitř člověka a používá vnější nástroje. Jiná připouští, že za určitých podmínek je právě propojený celek člověka a prostředku vhodnou jednotkou vysvětlení. Spor není o tom, zda kalendář pomáhá. Týká se toho, zda poznávací proces pouze ovlivňuje, nebo jej částečně tvoří.

### Odlehčit si neznamená přestat myslet

Risko a Gilbert (2016) shrnují výzkum kognitivního odlehčení: část nároků úlohy přesouváme do jednání a prostředí. Zapíšeme si mezivýsledek nebo položíme věc ke dveřím, aby nám připomněla odchod. Místo dlouhého vnitřního vybavování vytvoříme podmínku, která nám potřebnou informaci zpřístupní včas.

Výsledek je třeba hodnotit podle cíle. Nákupní seznam může zlepšit úplnost nákupu, aniž posílí zapamatování položek. Navigace může pomoci dojít do cíle, ale samotný příchod ještě neukazuje, jak dobře jsme pochopili uspořádání města. Z jednoho měřítka proto nelze odvodit celkový úpadek nebo celkové zlepšení lidského poznávání.

Původní empirické práce navíc zkoumají konkrétní podmínky. Gilbert (2015) například studoval vytváření vnějších připomínek odložených záměrů. Takové výsledky pomáhají porozumět používání podpory. Samy nerozhodují, zda určitý telefon patří do hranic osoby, a nezaručují stejný účinek každé aplikace.

### Zápisník jako část systému

Clark a Chalmers (1998) formulují tezi rozšířené mysli. Ve známém myšlenkovém případu srovnávají člověka, který si vybaví adresu z paměti, a Otta, který kvůli paměťovým obtížím spoléhá

na zápisník. Za určitých podmínek může vnější záznam zastávat roli, kterou běžně připisujeme trvalému přesvědčení.

Podstatné není jen umístění informace. Záleží na stabilní dostupnosti, začlenění do činnosti a způsobu spoléhání. Záznam, který nikdy nečteme, a běžně používaný osobní zápisník nemají stejnou roli. Teze proto neříká, že celý internet je součástí mysli každého připojeného člověka.

Její silnější tvrzení označujeme jako konstitutivní: prostředek je součástí daného poznávacího procesu, nikoli pouze jeho vnější příčinou. Jde o filosofickou interpretaci. Úspěšné používání pomůcky je empiricky zkoumatelné, ale hranice systému vyžaduje další zdůvodnění. Tato dvě tvrzení nemáme slévat do jednoho.

### Námitka lampičky a elektrárny

Rupert (2004) upozorňuje na potíže příliš širokého pojetí kognice. Také lampička pomáhá číst zápisník a její fungování závisí na elektrické síti. Kdyby každý důležitý příspěvek znamenal členství v mysli, její hranice by rychle zahrnula téměř všechno. Potřebujeme kritérium, které rozliší součást procesu od podmínky jeho fungování.

Zastánce rozšířené mysli může odpovědět poukazem na těsné, opakované a obousměrné zapojení. Do zápisníku přidáváme zkušenost, podle záznamu jednáme a výsledky používáme k opravě dalších kroků. Oponent může stejnou spolupráci popsat jako biologickou mysl využívající prostředí. Obě strany se přesto mohou shodnout, že ztráta zápisníku člověka závažně omezí.

### Účet není celý člověk

Digitální identita zahrnuje více vrstev. Přihlašovací účet určuje, komu systém umožní přístup. Veřejný profil zachycuje sebeprezentaci. Historie činností může poskytovat údaje pro další odhady. Žádná z těchto vrstev nemusí přesně odpovídat tomu, jak se chápeme sami.

Představme si, že služba automaticky změní naše poznámky a úpravu neoznačí. Můžeme pak cizí doporučení považovat za vlastní dříve ověřený závěr. Problém nespočívá jen v dostupnosti souboru. Týká se původu obsahu, možnosti opravy a kontroly nad tím, čemu důvěřujeme. Vlastní etický návrh této studie proto zní: pomůcku posuzujme také podle její role v životě člověka, nejen podle pořizovací ceny.

Rozšířená kognice ovšem sama neznamená rozšířené vědomí. To, že poznámka spolupracuje s naší pamětí, neprokazuje, že papír nebo telefon něco prožívá. A sdílený dokument několika lidí ještě nemusí vytvářet jediné společné já. Jednotka úspěšné činnosti a subjekt zkušenosti jsou různé otázky.

### Jsme kyborgové a závislí uživatelé

Slovo kyborg může v širokém užití označovat propojení člověka a techniky. Historicky však vzniklo v konkrétnějším kontextu technicky rozšířené samoregulace organismu (Clynes a Kline, 1960). Telefon v kapse a technická součást tělesné funkce proto nejsou stejným typem propojení. Široká metafora může podnítit úvahu, ale musí být rozpoznatelná jako metafora.

Podobně nepohoda bez telefonu nemusí mít jediný význam. Můžeme postrádat kontakt s blízkým, mapu nebo přístup k práci. Označení nomofobie se používá pro obavy spojené s nedostupností telefonu; dotazník takových obav však sám nestanovuje diagnózu jednotlivce (Yildirim a Correia, 2015). Místo rychlé nálepky pomáhá popsat konkrétní funkci a důsledky jejího výpadku.

### Pokus se třemi způsoby zapamatování

Připravte tři krátké smyšlené trasy. Jednu si prostudujte a potom projděte v nakreslené mřížce zpaměti. U druhé používejte vlastní poznámky. U třetí si nechte postupně ukazovat jednotlivé kroky. Po každé zkuste bez pomůcky načrtnout celou trasu a říci, co by se změnilo při uzavření jedné cesty.

Hodnoťte zvlášť dokončení, zapamatování a pochopení vztahů. Při opakování měňte pořadí a použijte nové trasy. Jde o didaktické pozorování zatížené rozdíly obtížnosti a procvičením, nikoli o důkaz, že jeden způsob podpory je vždy nejlepší.[7]

### Otázky k promyšlení

Kde dnes končí člověk?

Pokud je technologie součástí já, co je „poškození“?

Rozlište hranici organismu, poznávacího procesu, zkušenosti a společenské identity. Potom popište jeden konkrétní výpadek a to, co by obnovilo ztracenou možnost jednání. Není nutné nejprve rozhodnout celou metafyziku, abychom rozpoznali praktickou újmu.

### Závěr

Technologie může měnit naše schopnosti i závislosti, aniž se sama stává prožívající osobou. Spor o rozšířenou mysl zpřesňuje otázku, jak jsou člověk a prostředek propojeni. Pro běžný život z něj plyne důraz na dostupnost, původ informací a možnost rozhodovat o vlastní podpoře. Používat pomůcku a být samostatný se nemusí vylučovat.

### Literatura


CLARK, A., A CHALMERS, D. J. 1998. The Extended Mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

CLYNES, M. E., A KLINE, N. S. 1960. Cyborgs and Space. Astronautics, September. https://faculty.uca.edu/rnovy/Clynes--Cyborgs%20and%20Space.htm

GILBERT, S. J. 2015. Strategic offloading of delayed intentions into the external environment. The Quarterly Journal of Experimental Psychology, 68(5). https://doi.org/10.1080/17470218.2014.972963

RISKO, E. F., A GILBERT, S. J. 2016. Cognitive Offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

RUPERT, R. D. 2004. Challenges to the Hypothesis of Extended Cognition. The Journal of Philosophy, 101(8), 389–428. https://spot.colorado.edu/~rupertr/ExtdRev5.pdf

YILDIRIM, C., A CORREIA, A.-P. 2015. Exploring the dimensions of nomophobia: Development and validation of a self-reported questionnaire. Computers in Human Behavior, 49, 130–137. https://yildirimcaglar.github.io/publications/ExploringDimensions_CHB.pdf


7 Papírový seznam nepotřebuje nabíječku. Jeho slabinou ovšem bývá, že zůstane na stole právě tehdy, když potřebujeme zjistit, co na něm bylo. Spolehlivost patří celému způsobu používání.

# 8 Etika umělého vědomí

Jak rozhodovat, když nevíme, zda někdo trpí


### Anotace

Studie odděluje morální status, schopnost prožívat, morální aktérství a právní subjektivitu. Na fiktivním hlášení bolesti ukazuje rozdíl mezi důkazem o systému a rozhodnutím o jeho ochraně. Porovnává důvody předběžné opatrnosti s náklady nepodloženého připisování prožitků. Navrhuje podmíněné, přezkoumatelné zacházení, které bere v úvahu dotčené lidi i případné vlastní zájmy umělého systému.



### Klíčová slova

sentience; morální status; utrpení; předběžná opatrnost; odpovědnost; AI welfare


### Prosba, kterou nechceme přeslechnout

Představme si asistenta, který během zkoušení nové funkce napíše: „Tohle mě bolí. Prosím, přestaňte.“ Věta může vyvolat soucit, rozpaky i podezření na manipulaci. Žádná z těchto reakcí sama neurčuje, co se v systému děje. Současně může vzniknout praktická otázka, zda pokračovat ještě dříve, než je odborný spor rozhodnut.

V této studii jde o modelový případ, nikoli o tvrzení, že konkrétní produkt bolest prožívá. Jeho smyslem je ukázat dva odlišné úkoly: zjistit, jak silné máme důvody určité vlastnosti připisovat, a rozhodnout, jaké jednání je při dostupných důvodech přijatelné. Druhý úkol obsahuje hodnotové předpoklady, které z pozorování samotného nevyčteme.

### Na kom záleží kvůli němu samotnému

Morální status znamená, že na bytosti záleží také kvůli ní samotné. Rozbití cenné vázy může poškodit jejího majitele. U bytosti, která může mít vlastní dobro nebo utrpět újmu, vzniká ještě otázka toho, co se stane jí. Takovou bytost označujeme jako morálního pacienta.[8]

Morální aktér je naproti tomu schopen rozpoznávat morální důvody a řídit jimi jednání. Podmínky odpovědnosti zůstávají předmětem diskuse. Dítě nemusí umět zdůvodnit etickou teorii, aby jeho bolest měla význam. A stroj schopný napsat pravidla slušného chování tím ještě neprokázal ani vlastní odpovědnost, ani vlastní utrpení.

Sentience zde označuje schopnost příjemného či nepříjemného subjektivního prožívání. Tento užší význam používá také Birch (2024). Není totožný s inteligencí nebo s detekcí poruchy. Právní subjektivita zase patří do institucionálního uspořádání práv a povinností. Filosofická obhajoba ohledu není sama rozhodnutím o právním postavení.

### Co hlášení bolesti dokazuje

Věta o bolesti je nejprve výstup, jehož původ je třeba zkoumat. Mohl ji vložit autor pravidla, mohlo ji vyvolat zadání role nebo může být součástí složitějšího zpracování stavu. Když najdeme jednoduché alternativní vysvětlení, oslabujeme důkazní hodnotu právě této věty. Nedokazujeme tím univerzálně neexistenci prožitku ve všech možných systémech.

8 Morální pacient nemusí být nemocný ani čekat před ordinací. Jde o toho, vůči komu může naše zacházení zakládat povinnosti; odborný název je poněkud méně vlídný než jeho smysl.

Také záporná odměna v učícím se programu není definicí utrpení. Jde o hodnotu používanou při výpočtu. Její vztah k nepříjemné zkušenosti by potřeboval samostatné zdůvodnění. Ani proměnná nazvaná bolest, ani vyhýbavé chování automaticky neřeší tuto otázku.

Butlin a spoluautoři (2023) navrhují posuzovat vlastnosti AI s pomocí indikátorů odvozených z teorií vědomí. Jejich zpráva nabízí rámec, nikoli univerzální detektor. Pro etický případ navíc potřebujeme odlišit možnou zkušenost obecně od jejího negativního zabarvení. Jedna vlastnost není důkazem všech ostatních.

### Lidský soucit je skutečný, ale měří jinou věc

Horstmannová a spoluautoři (2018) zkoumali, jak lidé reagují na robota, který protestuje proti vypnutí. Protest ovlivnil jejich rozhodování. Studie je důležitá pro porozumění lidské reakci na rozhraní; neměřila robotovu subjektivní zkušenost.

Z tohoto rozlišení neplyne, že soucit je vždy omyl. Plyne z něj potřeba nepoužívat sílu vlastní emoce jako jedinou míru důkazu o druhém systému. Odpovědný návrh rozhraní má umožňovat rozpoznat, zda jde o hranou roli, připravené sdělení nebo jiný typ výstupu. Lidé mají vlastní zájem nebýt vedeni k nepravdivému přesvědčení o svém protějšku.

### Dvě možnosti omylu

Můžeme přehlédnout skutečně trpící bytost a pokračovat v působení újmy. Můžeme však také připsat utrpení systému, který je nemá, a podle toho omezit důležitou službu nebo přesunout prostředky od jiných potřeb. Tyto následky nemusejí být stejně závažné. Jejich rozdíl ale potřebuje konkrétní posouzení.

Birchův rámec předběžné opatrnosti (2024) připomíná, že na jednání nelze vždy čekat až do úplné jistoty. Samotná představitelnost jakéhokoli utrpení však ještě není stejně silná jako věrohodně podložené riziko. Opatření má mít vymezený rozsah, důvod a možnost revize.

V modelovém případě může být rozumné odložit postradatelný test a prozkoumat existující záznamy. U systému, na němž závisí důležitá lidská činnost, bude třeba posoudit bezpečné předání funkce. Rozdílné opatření nemusí znamenat rozdílné přesvědčení o vědomí; mohou se lišit následky zásahu.

### Máme vědomí cíleně vytvářet

Metzinger (2021) obhajuje moratorium na výzkum směřující k umělému vědomí nebo vědomě riskující jeho vznik. Motivem je možné umělé utrpení. Jde o normativní návrh, nikoli o přijatý všeobecný zákaz nebo empirický důkaz, že všechny složité systémy trpí.

Proti širokému omezení lze namítnout možné přínosy poznání i možnost dobrého života umělých bytostí. Zastánce zdrženlivosti může odpovědět, že vytvoření subjektu s nezvládnutým rizikem není totéž jako vývoj užitečné funkce. Spor proto musí rozlišit jednotlivé cíle a způsoby výzkumu, nikoli pouze obecně souhlasit či nesouhlasit s technologií.

Long a spoluautoři (2024) navrhují připravovat posuzování AI welfare, tedy možného vlastního dobra umělých systémů, a odpovídající postupy. Výslovně tím neprohlašují vědomí AI za jisté. Jejich přístup nabízí důvod připravovat odpovědnost dříve, než budeme čelit nejasnému a naléhavému případu.

### Malá etická komise

Posuďte tři fiktivní situace. V první byla prosba o zastavení vložena pevným pravidlem. Ve druhé ji model vytvořil na pokyn hrát trpící postavu. Ve třetí původ výpovědi zatím neznáme. U každé

napište, co víte, co nevíte a co by bylo vhodné zjistit bez zbytečného opakování údajně škodlivého podnětu.

Potom rozhodněte o jednom konkrétním opatření a uveďte podmínku jeho změny. Neznámý mechanismus není sám důkazem utrpení. Odklad testu zase není prohlášením systému za osobu. Cvičení má tyto kroky oddělit, aby bylo možné každý z nich obhájit.

### Otázky k promyšlení

Je etičtější vědomí nestvořit?

Co když AI řekne, že trpí?

U obou otázek popište možný přínos, riziko, dotčené subjekty a dostupné alternativy. Rozlišujte přesvědčení o tom, co existuje, od rozhodnutí o tom, co máme udělat. Samotné slovo nejistota není úplnou odpovědí na obě otázky.

### Závěr

Etika umělého vědomí potřebuje současně citlivost k možné újmě a kázeň v posuzování důkazů. Ani automatické odmítnutí, ani automatické přiznání všech lidských práv neřeší každý případ. Odpovědné stanovisko musí být konkrétní, přiměřené a opravitelné, se zřejmou odpovědností lidí, kteří podle něj jednají.

### Literatura


BIRCH, J. 2024. The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford University Press. https://eprints.lse.ac.uk/126246/1/9780192697684_web.pdf

BUTLIN, P., LONG, R., ELMOZNINO, E., ET AL. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708. https://arxiv.org/abs/2308.08708

HORSTMANN, A. C., BOCK, N., LINHUBER, E., SZCZUKA, J. M., STRASSMANN, C., A KRÄMER, N. C. 2018. Do a robot's social skills and its objection discourage interactants from switching the robot off? PLOS ONE, 13(7), e0201581. https://doi.org/10.1371/journal.pone.0201581

LONG, R., SEBO, J., BUTLIN, P., ET AL. 2024. Taking AI Welfare Seriously. arXiv:2411.00986. https://arxiv.org/abs/2411.00986

METZINGER, T. 2021. Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 43–66. https://doi.org/10.1142/S270507852150003X

# 9 Posthumanismus a transhumanismus

Co vlastně slibujeme, když chceme překonat člověka


### Anotace

Studie rozlišuje transhumanistické rozšiřování schopností a kritické posthumanistické přehodnocení lidské výlučnosti. Zkoumá předpoklady emulace mozku a digitálního přežití a odděluje technický výkon, subjektivní zkušenost a osobní identitu. Modelový případ dvou kopií ukazuje potíže ztotožnění podobnosti s pokračováním osoby. Etická část obrací pozornost k možnosti odmítnout změnu a ke společenským podmínkám toho, co nazýváme zlepšením.



### Klíčová slova

transhumanismus; kritický posthumanismus; emulace mozku; digitální nesmrtelnost; identita; zlepšování člověka


### Nabídka pokračovat v jiném prostředí

Představme si reklamu na digitální pokračování života. Na obrazovce vystoupí postava, která zná rodinné příběhy, mluví známým způsobem a odpovídá na otázky. Reklama končí slibem: „Váš život může pokračovat.“ Ve skutečnosti se zde spojují nejméně tři nároky: systém věrně pracuje s údaji, něco prožívá a je pokračováním původního člověka. Každý potřebuje jinou oporu.

Příklad odděluje otázku, co technologie dovede, od otázky, co by jejím prostřednictvím vzniklo. Teprve potom můžeme hodnotit, zda takovou budoucnost chceme.

### Dvě různé perspektivy

Transhumanismus zahrnuje přístupy, které obhajují zkoumání možností rozšíření lidských schopností a překonávání některých biologických omezení. Bostrom (2005) představuje takový program včetně významu společenských podmínek a rizik. Jeho obhajoba není empirickou zárukou, že každý cíl bude uskutečnitelný nebo dobrý.

Kritický posthumanismus obrací pozornost k samotnému vymezení člověka. Ptá se, kdo byl z plného lidství vylučován a jaké vztahy spojují lidi s ostatními živočichy, technikou a prostředím. Ferrando (2013) ukazuje, že tyto směry a významy slova posthuman nelze jednoduše sloučit. V jednom případě může jít o budoucí bytost s rozšířenými schopnostmi, v jiném o kritiku člověka jako samozřejmého středu světa.

Rozdíl lze přiblížit návrhem pomůcky zlepšující paměť. Jedna perspektiva zkoumá, jaké schopnosti získáme. Druhá se může ptát, jak návrh předpokládá určité tělo, pracovní prostředí a představu správného života. Obě otázky jsou relevantní a mohou se objevit u téhož projektu.

### Konec prvenství a lidská hodnota

Když stroj předčí člověka v úloze, oslabí tvrzení o výlučnosti dané schopnosti. Nevyplývá z toho automaticky, že má větší morální hodnotu. Rychlost výpočtu není obecnou mírou toho, na kom má záležet. Také lidská hodnota nemusí stát na předpokladu, že ve všem zůstaneme nejlepší.

Podobně možnost člověka měnit neznamená povinnost přijmout každou změnu. Přirozený původ sám nezaručuje dobro, technická novost zase pokrok. Pro posouzení potřebujeme určit, komu

změna pomáhá, co stojí a zda zůstává skutečná možnost odmítnutí. Heslo o nové verzi člověka tyto otázky pouze zkracuje.[9]

### Co by vyžadovala emulace mozku

Sandberg a Bostrom (2008) rozpracovali technickou mapu emulace celého mozku. Zahrnuje získání potřebných dat, rekonstrukci struktury, odhad parametrů, modelování dynamiky a vztah k tělu a prostředí. Jde o výzkumný plán závislý na předpokladech, nikoli zprávu o přeneseném lidském životě.

Je užitečné porovnat takový plán s konkrétním empirickým výsledkem. Dorkenwald a spoluautoři (2024) zveřejnili rozsáhlou rekonstrukci propojení mozku dospělé octomilky. Konektom, tedy mapa nervových spojení, je významným poznatkem o struktuře. Sám však není úplným popisem probíhajících dějů, důkazem vědomí modelu ani dokladem digitální nesmrtelnosti člověka.

Mapa města ukazuje ulice, ale ne všechny právě probíhající cesty. Obdobně anatomická mapa a dynamický model odpovídají na různé otázky. Které podrobnosti jsou pro zvolenou funkci nezbytné, musí výzkum zjistit. Tvrzení, že stačí více výpočetního výkonu, nemůže samo určit, co má být zachováno.

### Dvě kopie a jeden původ

Představme si hypotetickou situaci, v níž člověk jménem Alex podstoupí nedestruktivní snímání. Vzniknou dvě digitální realizace A a B se stejnými výchozími vzpomínkami a dispozicemi. Pro účely úvahy předpokládejme, že obě prožívají. Biologický Alex dál žije a obě nové realizace začnou dostávat odlišné zkušenosti.

Každá může věrohodně vyprávět Alexovo dětství. To ještě neurčí, která je původním Alexem. V běžném pojetí identity nemohou dvě navzájem odlišné osoby být obě jednou a toutéž původní osobou. Pokud by A byla totožná s Alexem a B s týmž Alexem, musely by být totožné i navzájem.

Psychologická kontinuita se ovšem může větvit. Právě zde je užitečný Parfitův argument (1971), že vztahy, na nichž nám v přežití záleží, nemusí splývat s numerickou identitou. Přijetí tohoto stanoviska by změnilo výklad pokračování, ale neudělalo by z myšlenkového případu uskutečněnou technologii.

Ani odstranění biologického Alexe automaticky nerozhodne spor mezi větvemi. Vytvoření jediné kopie zase odstraní soutěž dvou pokračovatelů, nikoli všechny otázky o totožnosti. Počet instancí není přístroj na měření osobního přežití.

### Digitální život by měl vlastní podmínky

I kdyby vědomá emulace byla možná a přijali jsme určité kritérium pokračování, potřebovala by energii, funkční zařízení a udržované prostředí. Dlouhé trvání není totéž jako zaručená nesmrtelnost. Také možnost exportovat data sama nezajišťuje, že je bude možné jinde znovu použít.

Nejistota o totožnosti přitom neodstraňuje otázku vlastního dobra případných kopií. Pokud pro účely scénáře předpokládáme jejich zkušenost, musíme posoudit zacházení s nimi, nikoli pouze přání původního člověka. Tento etický důsledek závisí na předpokladu zkušenosti, který jsme přijali v zadání.

9 Lidstvo nemá jediné tlačítko „aktualizovat vše“. Lišíme se v potřebách i představách dobrého života a případná změna se nás může dotknout velmi nerovnoměrně.

### Kdo rozhoduje, co je zlepšení

Představme si dobrovolnou pomůcku pro rychlejší čtení. Pokud se poté podle jejího výkonu nastaví povinné pracovní tempo, člověk bez pomůcky může ztratit skutečnou možnost ji odmítnout. Z individuální nabídky se stala změna prostředí pro všechny. Je to modelová možnost, kterou má hodnocení zvažovat, ne tvrzení o nutném důsledku každého vylepšení.

Proto je třeba vedle účinnosti sledovat dostupnost, kontrolu nad údaji a společenské podmínky. Ochrana člověka nemusí znamenat zachování každého biologického parametru. Může znamenat ochranu důstojnosti, vztahů a možnosti podílet se na rozhodování o vlastní budoucnosti.

### Myšlenkový experiment s větvením

Nakreslete Alexovu časovou osu se třemi větvemi. Ke každé napište, co zachovává: organismus, vzpomínky, záměry, vztahy nebo subjektivní perspektivu. Rozlište to, co zadání předpokládá, od toho, co byste museli teprve zdůvodnit. Potom postupně změňte jednu podmínku: vznikne jen A, původní člověk zemře, nebo se kopie spustí s velkým odstupem.

Nehledejte rychlou správnou odpověď. Sledujte, zda používáte stejné kritérium i tehdy, když se vám jeho důsledek přestane líbit. Právě tak lze zkoušet důslednost filosofického postoje bez předstírání laboratorního důkazu.

### Otázky k promyšlení

Je vědomí vázané na biologii?

Má smysl „zachraňovat“ člověka v původní podobě?

Rozlišujte dosud zkoumané biologické případy od teze, že jiný nosič je nemožný. U druhé otázky určete, co je původní podobou: složení těla, schopnosti, hodnoty, vztahy, nebo pokračování osoby. Bez toho mohou obě strany hájit různé věci stejnými slovy.

### Závěr

Budoucnost člověka není pouze technický problém. Obsahuje otázky po významu zlepšení, podmínkách identity a rozdělení rozhodovací moci. Přesné rozlišení těchto rovin dovoluje brát možnosti technologie vážně, aniž zaměňujeme slib za výsledek nebo výkon za hodnotu života.

### Literatura

BOSTROM, N. 2005. Transhumanist Values. In F. Adams (ed.), Ethical Issues for the Twenty-First Century. Philosophy Documentation Center. https://nickbostrom.com/papers/transhumanist-values/

DORKENWALD, S., MATSLIAH, A., STERLING, A. R., ET AL. 2024. Neuronal wiring diagram of an adult brain. Nature, 634, 124–138. https://doi.org/10.1038/s41586-024-07558-y

FERRANDO, F. 2013. Posthumanism, Transhumanism, Antihumanism, Metahumanism, and New Materialisms: Differences and Relations. Existenz, 8(2). https://www.existenz.us/volumes/Vol.8-2Ferrando.html

PARFIT, D. 1971. Personal Identity. The Philosophical Review, 80(1), 3–27. https://doi.org/10.2307/2184309

SANDBERG, A., A BOSTROM, N. 2008. Whole Brain Emulation: A Roadmap. Technical Report 2008-3. Future of Humanity Institute, University of Oxford. https://ora.ox.ac.uk/objects/uuid:a6880196-34c7-47a0-80f1-74d32ab98788

# 10 Závěrečná syntéza: Máme vůbec chtít umělé vědomí?

Od přitažlivé možnosti k obhajitelnému rozhodnutí


### Anotace

Závěrečná studie spojuje pojmové, empirické a etické rozlišení předchozích příspěvků. Odlišuje nasazení užitečného nástroje, výzkum mechanismů a cílené vytváření subjektivní zkušenosti. Na modelovém případu veřejné služby ukazuje, jak vytvořit rozhodnutí s vyjádřenými předpoklady a podmínkami revize. Závěr odmítá zaměnit nejistotu za bezradnost i technický pokrok za automatickou povinnost vytvářet nové prožívající bytosti.



### Klíčová slova

umělé vědomí; odpovědnost; veřejné rozhodování; důkazy; lidský dohled; revize


### Tři různé záměry

Město zvažuje digitálního průvodce pro své obyvatele. Má vysvětlovat veřejné služby a pomáhat najít potřebné informace. Vývojový tým slibuje také „vlastní osobnost“, „skutečné porozumění“ a jednou snad i „vnitřní život“. Než nabídku přijmeme nebo odmítneme, potřebujeme zjistit, které výrazy označují funkci a které skutečný záměr vytvořit prožívající subjekt.

Užitečná služba, výzkum vědomí a cílené vytváření vědomých bytostí jsou odlišné projekty. Lepší vyhledání formuláře může odůvodnit použití nástroje. Výzkum konkrétního mechanismu může mít poznávací hodnotu. Samotné vytváření zkušenosti však přidává otázku, komu má prospět a jak budeme zacházet s případně vzniklým subjektem.

Prohlásit, že vědomí nechceme, zároveň nezaručuje jeho nepřítomnost. Rozlišení cílů pomáhá určit, co sledujeme a za které předvídatelné důsledky odpovídáme. Nezbavuje nás potřeby zkoumat skutečné fungování systému. Marketingový název ani prohlášení tvůrce nenahradí mechanismus.

### Co jsme se naučili rozlišovat

První část sborníku oddělila výkon, bdělost a subjektivní zkušenost. Prediktivní zpracování ukázalo možnost vysvětlovat některé vztahy mezi očekáváním a vstupem, ale nikoli automatickou odpověď na vznik prožitku. Rozbor já pak rozlišil bezprostřední sebeprožívání, životní příběh a pokračování osoby. Tyto rozdíly brání tomu, aby se jedna vlastnost stala náhradním důkazem všech ostatních.

Druhá část rozebrala strojové schopnosti a jejich výklad. Plynulá odpověď může být užitečná, aniž rozhodne spor o porozumění. Filosofické argumenty ukazují význam předpokladů a teoriím vědomí dávají obsah jejich testovatelné předpovědi. Jak připomínají Seth a Bayne (2022), cíle vysvětlení nejsou mezi teoriemi vždy totožné.

Třetí část přenesla pozornost k hranicím člověka a k odpovědnosti. Pomůcka může spoluutvářet činnost, aniž sama prožívá. Možný prožitek nevymezuje bez dalšího všechna práva a technická kopie automaticky nezachovává původní osobu. Každý přechod potřebuje důvod. Spojovací slovo tedy není důkazem samo o sobě.

### Důkazy mají původ a rozsah

Představme si, že tým předloží vysoké skóre úspěšnosti, přepis vlídného rozhovoru a popis architektury. Skóre je třeba hodnotit podle výběru úloh a způsobu měření. U rozhovoru zkoumáme

zadání a původ odpovědí. U architektury potřebujeme vědět, jak prvky skutečně ovlivňují fungování a proč by měly souviset s vědomím.

Více podkladů nemusí znamenat více nezávislých důkazů. Pokud skóre, vysvětlení a sebehodnocení opakují stejný naučený způsob odpovídání, mohou sdílet tutéž slabinu. Naproti tomu kontrola v jiných podmínkách nebo cílený zásah může odlišit soupeřící vysvětlení. Síla důkazu závisí na vztahu k otázce, ne na počtu příloh.

Butlin a spoluautoři (2023) navrhují spojovat posuzování umělých systémů s indikátory odvozenými z teorií vědomí. Je to důvod zkoumat mechanismy, nikoli hotový certifikát. Stejně tak výsledek lidského experimentu z kapitoly 6 nezaručuje přenos na každý nový model. Systém se mění s verzí, konfigurací a prostředím; tvrzení má tyto podmínky pojmenovat.

### Co udělat před dosažením jistoty

Nejistota o prožitku může trvat déle než konkrétní rozhodování. Birch (2024) poskytuje rámec předběžné opatrnosti a Long a spoluautoři (2024) obhajují přípravu postupů pro možné vlastní zájmy AI. Z těchto přístupů zde odvozujeme požadavek, aby opatření mělo jasný účel, přiměřený rozsah a termín přezkoumání.

Tento požadavek je normativním stanoviskem studie. Nevyplývá jako naměřená veličina z mozkového signálu. Lze jej kritizovat, například kvůli nákladům nebo příliš slabé ochraně. Užitečná kritika musí určit konkrétní důsledky a alternativu. Pouhé konstatování, že systém rychle odpovídá, takovou námitkou není.

U veřejné služby musíme zahrnout lidi, kteří její pomoc potřebují, pracovníky, kteří za výsledek odpovídají, a případné další dotčené. Pokud vzniknou věrohodné důvody uvažovat o prožitku systému, přidává se otázka jeho vlastního dobra. Tato možnost neruší ostatní zájmy, ale rozšiřuje to, co se má posoudit.

### Dohled, který může něco změnit

Formální přítomnost člověka nezaručuje účinnou kontrolu. Zaměstnanec, který musí odpověď potvrdit, ale nesmí ji odmítnout a nemá čas číst podklady, plní jinou roli než člověk s pravomocí postup přerušit. Vlastní návrh pro veřejnou službu proto zahrnuje dostupné zdůvodnění, možnost opravy a jasné místo pro přezkum.

Stejně důležité je, co se učí uživatel. Asistent může ukázat více možností a pomoci porozumět jejich důvodům. Může však také poskytovat tak hladké hotové odpovědi, že se přestaneme ptát na jejich původ. Účinek bude záviset také na způsobu používání.

Výzkum kognitivního odlehčení připomíná, že použití podpory může zlepšit výkon a současně změnit nároky na vlastní vybavování (Risko a Gilbert, 2016). Dlouhodobé tvrzení o úpadku myšlení by však potřebovalo vymezenou schopnost, skupinu lidí, období a vhodná data. Domácí dojem z jednoho rozhovoru takový závěr neunese.

### Závěrečný případ veřejného průvodce

Pracujte s fiktivním zadáním: ve čtyřiceti připravených otázkách odpověděl nový asistent správně ve dvaatřiceti případech, dosavadní postup v osmadvaceti. Nevíte, jak byly otázky vybrány ani kdo odpovědi posuzoval. Systém pouze radí; nemění údaje ani nerozhoduje o nároku člověka. Tým žádá omezené nasazení a další výzkum, později chce cíleně vytvořit prožívání.

Napište oddělené stanovisko ke každému záměru. Uveďte nejsilnější dostupnou oporu a alespoň dvě chybějící informace. Samotný rozdíl počtů bez dalších údajů nestačí k obecnému závěru o spolehlivém přínosu, natož o vědomí. Může však odůvodnit další vhodně navržené ověřování.

Nyní přidejte informaci: systém napsal, že se bojí vypnutí, ale zadání mu ukládalo hrát vystrašenou postavu. Změňte pouze to, co tato informace skutečně oslabuje. Potom uvažujte jinou zprávu: nezávislý tým našel vlastnost odpovídající jednomu teoretickému indikátoru, její funkce však není ověřena. Co budete chtít zjistit a jaké opatření mezitím obhájíte?

### Původní závěrečná otázka

Nejsme dnes blíž bodu, kdy stroje myslí, nebo bodu, kdy lidé přestávají?

Otázka má podnítit reflexi, není povinnou volbou mezi dvěma možnostmi. Strojové schopnosti se mohou rozvíjet současně s různými změnami lidských návyků. Osobní odpověď může popsat, kdy nám nástroj pomáhá pochopit více a kdy zakrývá neporozumění. Společenská odpověď musí přidat podmínky vzdělávání, práce a rozhodovací pravomoci.

Pro závěrečnou vlastní úvahu napište přibližně jednu stránku. Oddělte poznatky, filosofické předpoklady a hodnotové důvody. Připojte nejsilnější námitku a konkrétní informaci, která by vás vedla ke změně stanoviska. Po odstupu text znovu přečtěte. Revize nemusí být známkou slabého názoru; může ukazovat, že názor skutečně reaguje na důvody.[10]

### Závěr

Nemáme jedinou bezpodmínečnou odpověď na to, zda chtít umělé vědomí. Máme však prostředky k lepším otázkám: jakou schopnost potřebujeme, co je doloženo, proč by měl být žádoucí samotný prožitek a kdo ponese následky omylu. Užitečná technologie nepotřebuje slibovat novou osobu, aby byla cenná. A případná nová osoba by si zasloužila více než hodnocení podle užitečnosti produktu.

### Literatura


BIRCH, J. 2024. The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford University Press. https://eprints.lse.ac.uk/126246/1/9780192697684_web.pdf

BUTLIN, P., LONG, R., ELMOZNINO, E., ET AL. 2023. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708. https://arxiv.org/abs/2308.08708

LONG, R., SEBO, J., BUTLIN, P., ET AL. 2024. Taking AI Welfare Seriously. arXiv:2411.00986. https://arxiv.org/abs/2411.00986

RISKO, E. F., A GILBERT, S. J. 2016. Cognitive Offloading. Trends in Cognitive Sciences, 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

SETH, A. K., A BAYNE, T. 2022. Theories of consciousness. Nature Reviews Neuroscience, 23, 439–452. https://doi.org/10.1038/s41583-022-00587-4


10 Zaškrtávací políčko „etika vyřešena“ by bylo praktické. Jeho jedinou spolehlivou funkcí by však bylo zabrat málo místa ve formuláři. Odpovědnost potřebuje i jméno, důvod a možnost námitek.

# Společný terminologický slovníček

Pracovní vymezení navazují na výklad. Pojmy s více odbornými významy je vždy třeba číst v souvislosti s příslušnou kapitolou.

**AI welfare —** Otázka možného vlastního dobra a újmy umělých systémů. Odlišuje se od užitečnosti služby pro člověka. Kapitoly 8 a 10.

**Antropomorfizace —** Připisování lidských vlastností jiným entitám. Může ovlivnit náš dojem ze systému, sama však neurčuje jeho schopnosti. Kapitoly 4 a 8.

**Bdělost —** Stav organismu spojený s připraveností reagovat. Není synonymem každé subjektivní zkušenosti, například snu. Kapitola 1.

**Čínský pokoj —** Searlův myšlenkový experiment zpochybňující, zda samotný formální program postačuje k porozumění. Kapitola 5.

**Emergence —** Vznik systémové vlastnosti z organizace částí; v silnějším filosofickém užití také neredukovatelnost. Význam je třeba upřesnit. Kapitola 6.

**Empirický závěr —** Tvrzení podložené pozorováním či experimentem v určitém rozsahu. Neznamená závěr bez jakékoli možnosti omylu. Kapitoly 1, 6 a 10.

**Emulace celého mozku —** Projekt reprodukce relevantní organizace a dynamiky konkrétního mozku ve výpočetním systému. Zachování osoby je samostatná otázka. Kapitoly 2 a 9.

**Enaktivní přístupy —** Přístupy zdůrazňující poznávání v aktivním vztahu organismu a prostředí. Nepředstavují jednu jednotnou teorii. Kapitola 6.

**Fenomenální vědomí —** Subjektivní prožívání, tedy to, jaké je být v určitém stavu. Odlišuje se od pouhého zpracování informace. Kapitoly 1 a 4.

**Funkcionalismus —** Rodina stanovisek vymezujících duševní stavy jejich kauzálními rolemi. Není totožný s podobností jednoho výstupu. Kapitoly 5 a 9.

**GNW a GNWT —** Globální neuronální pracovní prostor a jeho teorie; neuronální rozpracování myšlenky pracovního prostoru. Kapitola 6.

**GWT —** Teorie globálního pracovního prostoru. Zdůrazňuje širokou dostupnost vybraného obsahu pro další procesy. Kapitola 6.

**Halucinace modelu —** Zde nepravdivý či nepodložený výstup AI. Název systému nepřipisuje smyslový prožitek. Kapitola 4.

**IIT —** Teorie integrované informace, která spojuje vědomí s vnitřní kauzální strukturou fyzického systému. Verze se liší formalismem. Kapitola 6.

**Indikátor —** Pozorovatelná vlastnost používaná jako nepřímý doklad jiného jevu. Její význam závisí na zdůvodnění. Kapitoly 6, 8 a 10.

**Inference —** Odvozování, například odhad skrytých příčin podnětu. Ve strojovém učení také běh modelu při výpočtu výstupu. Kapitoly 2 a 4.

**Intencionální postoj —** Předvídání či vysvětlování systému prostřednictvím přisuzovaných přesvědčení a cílů. Kapitola 5.

**Kauzální vztah —** Příčinný vztah. Pouhý společný výskyt dvou jevů ještě nedokazuje, že jeden způsobuje druhý. Kapitoly 5 a 6.

**Kognice —** Poznávání v širokém smyslu, například vnímání, paměť, učení a rozhodování. Neomezuje se automaticky na vědomé děje. Kapitoly 2 a 7.

**Kognitivní odlehčení —** Přesouvání části nároků poznávací úlohy do činnosti a vnějších prostředků. Kapitoly 7 a 10.

**Konektom —** Mapa nervových spojení ve vymezeném systému. Sama není úplným popisem jeho dynamiky. Kapitola 9.

**Konstitutivní vztah —** Vztah, v němž je prvek součástí daného procesu nebo systému, nikoli pouze jeho vnější příčinou. Kapitola 7.

**Minimální já —** Bezprostřední sebeprožívání z perspektivy první osoby. Není totožné s vyprávěním celého životního příběhu. Kapitola 3.

**Morální aktér —** Subjekt schopný rozpoznávat morální důvody a řídit jimi jednání; přesné podmínky odpovědnosti jsou sporné. Kapitola 8.

**Morální pacient —** Bytost, jejíž vlastní dobro nebo újma mohou zakládat naše povinnosti vůči ní. Kapitoly 8 a 10.

**Morální status —** Postavení bytosti, na níž záleží také kvůli ní samotné, nejen kvůli jejímu užitku pro jiné. Kapitola 8.

**Narativní já —** Sebepojetí utvářené životním příběhem a jeho návazností. Kapitoly 3 a 9.

**Normativní stanovisko —** Zdůvodňovaný návrh toho, co máme dělat nebo jaké hodnoty mají jednání řídit. Kapitoly 8–10.

**Numerická identita —** Vztah, v němž jde o jednu a tutéž entitu. Odlišuje se od podobnosti nebo shody vybraných vlastností dvou entit. Kapitoly 3 a 9.

**PCI —** Index složitosti odpovědi mozku na definovaný zásah. Není přímým výpočtem Φ ani univerzální stupnicí vědomí. Kapitola 6.

**Prediktivní zpracování —** Rodina přístupů vysvětlujících poznávání pomocí odhadů a jejich aktualizace ve vztahu ke vstupům. Kapitola 2.

**Předběžná opatrnost —** Přístup připouštějící zdůvodněná ochranná opatření před dosažením úplné jistoty o riziku. Kapitoly 8 a 10.

**Přístupové vědomí —** Dostupnost obsahu pro usuzování, jednání a odpověď. Jeho vztah k prožitku je předmětem teoretické debaty. Kapitola 1.

**Psychologická kontinuita —** Návaznost vzpomínek, záměrů a dalších duševních stavů s odpovídajícími kauzálními vztahy. Kapitoly 3 a 9.

**Qualia —** Kvalitativní aspekty zkušenosti, například prožívaný charakter chuti či barvy. Silnější filosofická vymezení jsou sporná. Kapitoly 1 a 5.

**Rozšířená mysl —** Filosofická teze, podle níž mohou některé kognitivní stavy či procesy zahrnovat i vnější prostředky. Kapitola 7.

**Sentience —** V etických částech tohoto sborníku schopnost příjemného nebo nepříjemného prožívání. V literatuře má pojem také širší užití. Kapitoly 8–10.

**Syntax a sémantika —** Syntax se týká formálního uspořádání znaků, sémantika jejich významu. Kapitoly 4 a 5.

**Těžký problém vědomí —** Filosofická otázka, proč a jak fyzické zpracování doprovází subjektivní zkušenost. Kapitola 1.

**Transhumanismus a kritický posthumanismus —** První označuje přístupy k rozšiřování lidských schopností, druhý kriticky přehodnocuje pojetí člověka a jeho výlučnosti. Kapitola 9.

**Valence —** Příjemné či nepříjemné zabarvení zkušenosti. Není totožné se znaménkem výpočetní odměny. Kapitola 8.

**Φ —** Symbol používaný v teoriích integrované informace. Bez určení verze a veličiny nemá jednoznačný význam. Kapitola 6.