Vyhodnocování inteligence umělých systémů

4IZ431: Přednáška 2

Ondřej Vadinský

KIZI
VŠE Praha

ZS 2021

Obsah

Úvod

- **Mohou stroje myslet a jak to poznat?**
- *Filosofické a kognitivní předpoklady inteligence.*
- *Přehled přístupů vymezující, definujících a testujících inteligenci umělých systémů.*
- Definice umělé inteligence: Wang (2019).

Slabá a specifická umělá inteligence

Searle (1980) – slabá UI:

- UI jako nástroj pro poznání lidské mysli,
- program jako přesnější formulace hypotézy o mysli,
- program umožňuje **simulovat mentální schopnosti**.

Goertzel (2014) – specifická UI:

- UI jako tvorba programů pro **řešení specifických úloh**,
- (šachové programy, i sofistikované jako osobní asistentka Siri).

Silná a obecná umělá inteligence

Searle (1980) – silná UI:

- UI (počítač vykonávající správný program) má rozumění a další kognitivní stavy,
- program jako vysvětlení mysli,
- UI je snaha vytvořit **stroj srovnatelný s člověkem**.

Goertzel (2014) – obecná UI:

- UI jako tvorba programů pro **obecné řešení úloh** a **obecné inteligentní jednání**,
- inteligence je obecná schopnost podkládající jiné schopnosti mysli.

René Descartes

- **Metodologický skepticizmus**:
  - Skrze pochybování k pevným principům vědění.
  - *„Myslím, tedy jsem.“*
- **Racionalizmus** × empirizmus:
  - *rozum* jako rozhodující *zdroj vědění*,
  - *smysly klamou*.
- **Dualizmus**:
  - nemateriální *mysl*,
  - materiální *tělo a svět*,
  - dobrotivý Bůh zajišťuje shodu při vnímání.

Descartovo rozlišení člověka a stroje I

- *Dva předpoklady inteligence*:
  - **schopnost rozumné řeči**,
  - univerzálnost myšlení.

Descartes (1637), str. 41:

„...kdyby existovaly stroje, podobající se našim tělům a napodobující naše úkony [...] měli bychom vždy dva velice vážné důvody, abychom poznali, že proto ještě nejsou skutečnými lidmi. První důvod je, že by nikdy nemohly užívat slov ani jiných znaků, skládajíce je jako činíme my, abychom své myšlenky vyložili jiným. Neboť lze dobře chápat, že stroj může býti udělán tak, aby pronášel slova, ba dokonce aby pronášel některá ve spojení s tělesnými úkony, souvisejícími s nějakými změnami jeho orgánů: jako například když se ho dotkneme na určitém místě, aby se zeptal, co mu chceme říci, když na jiném místě, aby křičel, že ho to bolí, a podobně; nemůže však být udělán tak, aby slova různě sestavoval a takto odpovídal na vše, co se řekne v jeho přítomnosti, jak to i nejtupější lidé mohou činit.“

Descartovo rozlišení člověka a stroje II

- *Dva předpoklady inteligence*:
  - schopnost rozumné řeči,
  - **univerzálnost myšlení**.

Descartes (1637), str. 41:

„...kdyby existovaly stroje, podobající se našim tělům a napodobující naše úkony [...] měli bychom vždy dva velice vážné důvody, abychom poznali, že proto ještě nejsou skutečnými lidmi. [...] A druhý důvod je, že i kdyby vykonávaly určité věci stejně dobře nebo snad i lépe než kdokoli z nás, selhaly by nevyhnutelně v jiných, při nichž by vyšlo najevo, že nejednaly s vědomím, nýbrž toliko podle sestavení svých orgánů; neboť rozum je všestranný nástroj, kterého lze užívat ve všech možných případech, kdežto tyto orgány musí mít nějaké zvláštní uzpůsobení pro každý úkon jednotlivý, a proto je morálně nemožné, aby rozmanitost těchto orgánů v jednom stroji stačila přivést jej k tomu, aby jednal za všech okolností života stejně, jako jednáme my vlivem svého rozumu.“

„Turingův obrat“

Turing (1950):

- **Mohou stroje myslet?**
  - Co je stroj?
  - Co je myslet?
- **Dokáže člověk rozlišit, zda komunikuje s člověkem nebo se strojem?**

Imitační hra I

- *Účastníci*:
  - rozhodčí,
  - muž,
  - žena.
- *Podmínky*:
  - fyzické oddělení účastníků,
  - nepřímá komunikace.
- *Cíl*:
  - **Rozhodnout** pomocí otázek, **kdo je muž** a **kdo žena** (rozhodčí).
  - Zmást rozhodčího (muž).
  - Pomoci rozhodčímu (žena).

Imitační hra II

- *Účastníci*:
  - rozhodčí,
  - člověk,
  - **stroj**.
- *Podmínky*:
  - fyzické oddělení účastníků,
  - **nepřímá komunikace**.
- *Cíl*:
  - **Rozhodnout, kdo je člověk a kdo stroj**:
    - Bude se rozhodčí mýlit stejně často jako v předchozím případě?
    - Bude při 5minutovém rozhovoru úspěšnost běžného rozhodčího nižší než 70 %?

Stroje v imitační hře

- Digitální počítač (**Turingův stroj**):
  - paměť,
  - výpočetní jednotka,
  - program.
- **Univerzálnost digitálních počítačů**:
  - Turingovy stroje jsou deterministické.
  - Digitální počítač může předpovědět budoucí stavy a tak napodobit libovolný Turingův stroj.
  - Namísto stavby specifického stroje stačí napsat specifický program a ten spustit na libovolném počítači.

Námitky k imitační hře I

Teologická námitka

- *Schopnost myslet vychází z nesmrtelné duše, kterou dal Bůh člověku, ale ne stroji.*
- Turing:
  - Jde proti všemocnosti Boha: Proč by nemohl dát duši stroji?
  - Teologické argumenty bývají nespolehlivé.

„Strkání hlavy do písku“

- *Myslící stroje? To by bylo strašné! Stroje tedy nemyslí.*
- Turing:
  - Vychází z přesvědčení o nutnosti nadřazenosti člověka.
  - Nejde o solidní argument.

Námitky k imitační hře II

Matematická námitka

- *Existují dokázaná omezení strojů, která neplatí pro lidský rozum:*
  - **Gödelovy věty o neúplnosti** – existují logické formule, které nejsou dokazatelné ani vyvratitelné.
  - **Nerozhodnutelnost problému zastavení** – neexistuje program, který by pro všechny dvojice program – vstup rozhodl, zda se daný program zastaví.
- Turing:
  - Nebylo dokázáno, že se omezení netýkají i lidského rozumu.
  - Stroje se tedy mohou v odpovědích mýlit, stejně jako lidé.

Námitky k imitační hře III

Protiargument vědomí

- *Myšlení spjaté s emocemi, rozuměním a vědomím, což stroje nemají.*
- Turing:
  - V silném pojetí **solipsismus** – mohu tvrdit s jistotou jen o sobě.
  - Ve slabém pojetí – do TT lze zakomponovat **viva voce metodu** (rozhovor ukáže, zda se student látku naučil zpaměti, nebo jí rozumí).

Protiargument „neschopností“

- *Stroje nikdy nebudou schopny dělat X.*
- Turing:
  - Indukcí z nedostatečné dosavadní zkušenosti.
  - *...dělat chyby* – lze naprogramovat, *...být předmětem svého myšlení* – stroj může upravit své procedury (zpětnovazební a genetické algoritmy).

Námitky k imitační hře IV

Námitka Lady Lovelace (k Babbageovu analytickému stroji)

- *Analytický stroj není schopen tvořivého myšlení, dělá jen to, k čemu je naprogramován (co programátor ví, jak udělat).*
- *Stroj nás nemůže překvapit.*
- Turing:
  - Nevyplývá, že nelze naprogramovat myšlení a učení.
  - Stroje nás překvapují, i když je to často chybou v našem očekávání.

Protiargument spojitosti nervového systému

- *Spojitou CNS nelze napodobit diskrétním digitálním počítačem.*
- Turing:
  - Digitální počítač je schopen **imitovat spojitý počítač** dost přesně, aby je rozhodčí nerozeznal.

Námitky k imitační hře V

Neformálnost lidského chování

- *Lidské chování není řízeno sadou formálních pravidel, člověk není stroj.*
- Turing:
  - **Pravidla chování** (rules of conduct) × **zákonitosti chování** (laws of behaviour).
  - Nejsme schopni poznat, zda známe úplnou množinu pravidel.

Námitka extrasenzorického vnímání

- *Telepat by mohl překonat bariéru vnímání, na které je test založen.*
- Turing:
  - Fenomén ač vědecky nepochopený je **statisticky významný**.
  - Zpřísnění podmínek testu („telepathy-proof room“).

Učící se stroje

- *Další reakce na Lady Lovelace*:
  - tvůrčí myšlení jako zásah nadkritického množství látky neutronem,
  - **cibulová metafora** pro mechanické vysvětlení mysli.
- Největším nedostatkem strojů je **absence správného programu**:
  - program pro **„dětskou mysl“**,
  - **učením** a shromažďováním zkušeností rozvoj na úroveň **„dospělé mysli“**,
  - řízená evoluce programů pro „dětskou mysl“,
  - skrze učící se stroje lze překonat námitku Lady Lovelace.

Možné důsledky úspěchu ve hře

- **Stroj je srovnatelný s člověkem**:
  - má mysl,
  - má vědomí.
- *Stroj je lepší než neúspěšný hráč.*
- Odpověď na otázku, zda mohou stroje myslet.

Kritika Turingova testu

- Searle:
  - jazyk a rozumění,
  - intencionalita,
- Dennett:
  - jazyk a svět,
- Harnad:
  - plný repertoár lidského inteligentního chování,
- Schweizer:
  - externalizmus,
  - inteligentní chování druhu.

Silná a slabá umělá inteligence

Searle (1980):

- **Slabá umělá inteligence**:
  - UI jako *nástroj pro poznání lidské mysli*,
  - program jako *přesnější formulace hypotézy o mysli*.
- **Silná umělá inteligence**:
  - UI (počítač vykonávající správný program) *má rozumění a další kognitivní stavy*,
  - **program jako vysvětlení mysli**,
  - v podobě *symbolické hypotézy*.

Myšlenkový experiment s čínským pokojem I

Ověření teorie silné UI:

- **Jaké by to bylo, kdyby mysl fungovala jako program?**
- *Jaké to je být strojem?*

Předpoklady:

- Searle vůbec neumí čínsky.
- Searle je rodilý anglický mluvčí.

Myšlenkový experiment s čínským pokojem II

Searle zamknutý v pokoji:

- **čínské texty**:
  - *skript* – znalostní reprezentace,
  - *příběh*,
  - *otázky*.
- **anglické instrukce**:
  - *jak propojit znaky z čínských textů* (na základě jejich tvaru),
  - *jak k sadě znaků* (otázka) *sestavit jinou sadu znaků* (odpověď).
- **anglické texty**:
  - *příběh*,
  - *otázky*.

Myšlenkový experiment s čínským pokojem III

Rozumění a čínský pokoj:

- **Zvenku** – Searle v místnosti je instanciací programu pro čínštinu:
  - *Odpovědi v čínštině* jsou *nerozlišitelné od odpovědí* rodilého *Číňana*.
  - *Odpovědi v čínštině* jsou *stejně dobré* jako *odpovědi v angličtině*.
- **Zevnitř**:
  - **Searle čínštině nerozumí**, jen manipuluje s *neinterpretovanými* symboly.

Myšlenkový experiment s čínským pokojem IV

Důsledky pro silnou umělou inteligenci:

- **Počítače nerozumějí příběhům v čínštině ani jiném jazyce**:
  - protože *Searle jako počítač nerozumí čínsky*,
  - a protože počítače v jiných situacích *nemají nic navíc oproti Searlovi* v pokoji.
- Počítač se správným programem **není postačující ani nutnou podmínkou rozumění**:
  - protože fungující počítač s programem nerozumí,
  - protože *člověk dělající to, co počítač, nezíská rozumění.*
- *Čínský pokoj ukazuje nedostatečnost Turingova testu.*

Námitky proti čínskému pokoji I

Námitka systému

- *Searle možná nerozumí čínštině, ale systém jako celek rozumí.*
- Searle:
  - Searle je **jediné místo** v systému **schopné rozumění**, a pokud on nerozumí, ani systém nerozumí *(homunkulární strategie)*.
  - Searle **se naučí knihy** se znaky a pravidly **zpaměti** a manipulace bude provádět v hlavě, aniž by rozuměl tomu, co dělá *(internalizační strategie)*.

Robotická námitka

- *Počítač v robotickém těle bude moci vnímat a jednat a tak rozumět.*
- Searle:
  - Nestačí manipulace se symboly, potřebuje i **kauzální vztah ke světu**.
  - Jen **přidá další čínské symboly** na vstupu (vnímání) a výstupu (jednání) a víc instrukcí (program), ale žádné rozumění *(homunkulární strategie)*.

Námitky proti čínskému pokoji II

Námitka simulátoru mozku

- *Program simulující aktivace neuronů Číňana by měl rozumět čínštině.*
- Searle:
  - Přiznává, že **funkcionalizmus nestačí**.
  - Muž obsluhující ventily potrubí by čínsky stále nerozuměl a v potrubí a ventilech žádné rozumění není. *(homunkulární strategie)*
  - Simuluje jen formální vlastnosti mozku, ne jeho kauzální schopnosti.

Kombinovaná námitka

- *Robot řízený simulátorem mozku a chápaný jako systém.*
- Searle:
  - **Přisouzení rozumění** na základě **vnější podobnosti a chování**.
  - Protože **ale uvnitř** jsou jen formální manipulace s neinterpretovanými symboly, není tam **žádné rozumění**.

Námitky proti čínskému pokoji III

Námitka jiných myslí

- *Rozumění ostatních přisuzujeme jen na základě chování, splní-li počítač behaviorální test, musíme mu také připsat rozumění.*
- Searle:
  - Nejde o to, jak vím, ale co přisoudím. To ale musí být víc než jen výpočet.

Námitka mnoha domů

- *Analogové a digitální počítače jsou jen současný stav techniky, v budoucnu půjde sestrojit zařízení s potřebnými kauzálními schopnostmi pro vznik rozumění.*
- Searle:
  - Příliš se vzdaluje původnímu vymezení silné UI.
  - Nejde o testovatelnou hypotézu.

Intencionalita

Searle (1980):

- Pro inteligenci je potřeba *rozumění dané intencionalitou.*
- **Intencionalita**: vlastnost mentálních stavů, která je činí *jsoucí o*, či *zaměřené na věci a situace ve světě.*

Pstružina (2003) – tři aspekty intencionality:

- **zaměřenost** na jsoucno pro sebe,
- **zvýznamnění** pojmu vřazením do struktury pojmů a endoceptů,
- **reflexe** zaměřenosti a zvýznamnění.

Připisování rozumění a intencionality

- **Skutečné rozumění**:
  - podložené *intencionalitou* a *kauzálními schopnostmi mozku*,
  - může být částečné, neúplné.
- **Metaforické rozumění**:
  - připisované nástrojům, kterými *rozšiřujeme svou intencionalitu*,
  - připisované artefaktům či zvířatům *pouze na základě chování*.

Původ intencionality

- *Proč lidé rozumějí?*
  - *ne kvůli komputačním procesům nad formálními prvky*,
  - ale kvůli určitě **biologické** (chemické a fyzikální) **struktuře organizmu**, která dokáže za jistých podmínek **způsobit intencionální fenomény**.
- *Mohou stroje myslet, rozumět a mít intencionalitu?*
  - **ano**, pokud budou **duplikovat potřebné kauzální schopnosti**,
  - **ne**, pokud budou mít **jen komputační procesy** nad formálními prvky (programy).
- *Původ intencionality*:
  - v systému samotném (člověk),
  - z jiného systému (stroj – *intencionalita od programátora a uživatele*).

Selhání funkcionalizmu

- **Vztah mysl – mozek není jako vztah program – hardware**:
  - Program lze realizovat *„šílenými způsoby“*, kt. zjevně nemají intencionalitu (překládání kamínků).
  - **Program** je čistě formální (**jen syntaxe**), **intencionální stavy** jsou i o obsahu (**sémantika**).
  - Mentální stavy jsou produktem fungování mozku, program není produktem počítače.
- **Simulace není duplikací**:
  - **Počítače nezpracovávají informace**, ale **manipulují s neinterpretovanými symboly** (interpretace z vnějšku).
  - *Reziduální behaviorizmus v UI* (Turingův test).
  - *Reziduální dualizmus v UI* (programy jsou zcela nezávislé na hardwaru).

Problémy týkající se mysli a těla

Problém mysli a těla – viz Havel (2001)

- Metafyzický problém: **Zda a jaký je vztah mezi mentálními a tělesnými ději?**
  - *dualizmus* – substanční, vlastností, módů bytí,
  - *monizmus* – mentalizmus, idealizmus, materializmus, emergentizmus.

Problém jiných myslí – viz Hyslop (2014)

- Epistemologický a konceptuální problém: **Mají i jiní lidé mysl?**
  - *solipsizmus*,
  - *inference z analogie*,
  - *teorie inferovaná z chování*,
  - *chování jako kritérium přítomnosti mysli*.

Úplný Turingův test I

Harnad (1991):

- *Otázky jako:*
  - „Jsou stroje skutečně inteligentní?“
  - „Rozumějí skutečně stroje tomu, co dělají?“
  - „Mohou stroje skutečně vidět?“
- *jsou vlastně otázkou:*
  - **„Mají stroje mysl?“**
- **Problém jiných myslí** a jeho řešení je **relevantní pro test UI**:
  - Lidé přisoudí mysl jiným na základě *jejich chování v reálném světě*, které je nerozlišitelné o jejich vlastního chování.
  - *Robot musí v reálném světě* (s lidmi a objekty) *dokázat udělat vše, co běžní lidé*, pro běžného člověka nerozlišitelně od nich.

Úplný Turingův test II

- *Účastníci*:
  - **robot**,
  - člověk.
- *Podmínky*:
  - **přímá interakce se světem**,
  - **přímá komunikace**.
- *Cíl*:
  - Ověřit, **zda je robot schopen plného spektra běžného lidského chování**.
  - Ověřit, **zda robot rozumí jazyku**:
    - *jazykový vstup – behaviorální výstup*,
    - *behaviorální vstup – jazykový výstup*.

Úplný Turingův test a čínský pokoj

- *Úplný Turingův test*:
  - řeší vztahy uvnitř jazyka (jako Turingův test),
  - navíc **umožňuje propojit jazyk mimojazykovými vztahy se světem**.

Schweizer (2012) – Roboticko-systémová námitka:

- *Searlova homunkulární strategie není přesvědčivá*:
  - pro test **relevantní vstupy** a **výstupy nejsou** homunkulovi **přímo přístupné**,
  - homunkulus přebírá a předává symboly dalším subsystémům.
- *Searlova internalizační strategie není adekvátní*:
  - Searle **internalizuje** robota do **celého svého těla ne jen do vědomí**.

Externalizmus

Kripke (1972), Putnam (1975) a Burge (1979):

- *Interní reprezentace není dostatečná pro jazykovou referenci*:
  - **přímé kauzální vztahy s prostředím při osvojování jazyka** (intersubjektivně přístupný význam v prostředí),
  - **dělba lingvistické práce** (experti pro přesné vymezení významu).
- Jazyk je společenský, historicky evolučně vzniklý fenomén.
- **Význam** je veřejná záležitost daná:
  - *pravidelnostmi v mikrostruktuře*,
  - *kauzálními vazbami*,
  - *relevantními experty*,
  - *a zažitou praxí v sociolingvistickém klanu.*

Skutečně úplný Turingův test I

Schweizer (2012):

- Východisko: sémantický a mentální *externalizmus*,
- Lidská **mysl utvářená externími fenomény**:
  - jazyk,
  - kultura,
  - společnost.
- Tyto fenomény jsou **produktem druhu, ne jednotlivce** („type“ × „token“).
- **Testy UI zaměřené na druh**:
  - jednotlivci v rámci svého druhu,
  - ne jednotlivec v rámci jiného druhu.

Skutečně úplný Turingův test II

- *Účastníci*:
  - **roboti** (umělý druh),
  - lidé (biologický druh).
- *Podmínky*:
  - přímá interakce se světem,
  - **evoluce komunity**.
- *Cíl*:
  - Ověřit **schopnost robotů** (druhu) **vytvořit si vlastní inteligentní chování**,
  - Ověřit, **zda si roboti vyvinou jazyk**.

Skutečně úplný Turingův test a čínský pokoj

- **Čínský pokoj**:
  - **mysli mají mentální obsahy** (sémantiku),
  - výpočetní syntaxe není nutnou ani postačující podmínkou pro sémantiku.

Schweizer (2012) – skutečně úplný Turingův test

- zachovává důraz na **esenciálnost významu pro mentální obsahy**,
- nahrazuje interní sémantiku **externí sémantikou**,
- pro s lidmi srovnatelné rozumění jazyku pak *nestačí jen test jedince.*

Inteligence a kognitivní schopnosti

de Mey (1992):

- **Kognitivní paradigma** – zpracování informací vyžaduje model světa (reprezentaci).
- **Stratifikovaný model vnímání** spojuje:
  - *vjemy objektu*,
  - *očekávání subjektu*.
- **Získávání znalostí** za pomoci vnímání a jednání:
  - jednání nejprve formuje *implicitní znalosti*,
  - opakováním jednání a vnímáním následků dochází k *zexplicitnění implicitního*.

Hierarchie úrovní

Sun (2007):

- Různé disciplíny zkoumají kognici na *různých úrovních abstrakce*.
- Vysvětlení kognice pomocí **integrované hierarchie úrovní**:
  - **sociální** – společnost, kultura, organizace, prostředí, ...
  - **psychologické** – individuální chování, zkušenost, dovednosti, ...
  - **komponentové** – kognitivní mechanizmy a je realizující komponenty,
  - **fyziologické** – realizace v substrátu.
- Implementace úrovní ovlivněna:
  - *omezeními zdola*,
  - *omezeními shora*.

Kognitivní architektury I

Sun (2007) – kognitivní architektura:

- **doménově obecný** výpočetní kognitivní **model**,
- zachycující **podstatné struktury** a **procesy v mysli**,
- využitelný pro širokou, mnohoúrovňovou a mnohodoménovou analýzu chování.

- Použití:
  - **kognitivní věda** – framework pro modelování kognice:
    - nové poznatky o lidské kognici,
    - výzkum kolektivního lidského chování,
    - kognitivní realizmus.
  - **umělá inteligence** – infrastruktura inteligentních systémů.

Kognitivní architektury II

Sun (2007) – schopnosti a omezení kognitivních architektur:

- **Možné funkcionality**:
  - vnímání a kategorizace,
  - reprezentace, paměť a odvozování,
  - rozhodování a řešení problémů,
  - plánování a jednání,
  - komunikace,
  - metakognice, motivace a cíle,
  - učení.
- **Integrální × sekundární schopnosti**.

CLARION

Sun (2007) – kognitivní architektura CLARION:

- *Modulární architektura* (vždy hybridní systém):
  - **akční modul** – řízení vnějších a vnitřních akcí,
  - **neakční modul** – udržování znalostí,
  - **motivační modul** – motivace pro vnímání, jednání a kognici,
  - **metakognitivní modul** – sledování, řízení a úprava ostatních modulů.
- *Požadavky*:
  - učení bez apriorních doménově specifických znalostí,
  - kontinuální učení z interakce se světem,
  - odlišené typy znalostí, včetně specifických metod učení,
  - reaktivní i kognitivní pohled,
  - komplexní situace.

ACT-R

Anderson et al. (2004) – kognitivní architektura ACT-R:

- Základní principy:
  - **racionální analýza** – optimalizace komponent vůči prostředí,
  - nereduktivní rozdělení **znalostí** na **deklarativní** a **procedurální**.
  - **modulární struktura** s komunikací přes zásobníky
- *Skupiny modulů* (různé symbolické systémy):
  - **perceptuálně-motorické** – interakce se světem,
  - **paměťové moduly**:
    - **deklarativní paměť** – fakta jako „chunky“,
    - **procedurální paměť** – „productions“ jak reagovat na situaci.
- *Použití*:
  - paměť, pozornost, přirozený jazyk a komplexní úlohy,
  - neurověda (predikce mozkové aktivity), výuka („cognitive tutors“).

Uroboros

Thomsen (2013) – kognitivní model Uroboros:

- Vysvětlení kognice a vědomí přirozených a umělých agentů.
- **Algoritmický rekurzivní proces**:
  - *očekáváním řízené získávání dat* (vnímání a jednání),
  - *konzumační analýza* (vyhodnocování dat vzhledem k očekávání a aktualizace datových struktur).
- **Vhodné datové struktury**:
  - *nestriktní hierarchie schémat, „chunků“, rámců,*
  - speciální případy: *perfektní shoda* × *závažné nenaplnění očekávání*,
  - současné aktivace ukládány jako nové rámce.
- **Interpolace dat mezi časovými snímky**.

Psychometrická umělá inteligence (PAI)

Bringsjord – Schimanski (2003):

- využívá **psychometrické definice inteligence**.
- Psychometrie – **měření inteligence** a jiných mentálních schopnosti lidí za použití testů.
- UI by se měla zaměřit na "vytváření takových entit zpracovávajících informace, které budou schopné dosáhnout alespoň **solidních výsledků ve všech zavedených a ověřených testech inteligence a mentálních schopností.**."

Besold et al. (2015):

- **Přímé použití testů lidské inteligence je problematické** (postačující i nutné podmínky).
- Testy potřebují zobecnit a vylepšit.

C-Test

Hernandez-Orallo (2000):

- **Nedostatky přístupů typu Turingova testu**:
  - Jedná se spíš o myšlenkový experiment, než o praktický test.
  - Inteligence není absolutní atribut (nestačí ano × ne odpověď)
  - Inteligence je vícedimenzionální.
  - Test nesmí být antropomorfní.
  - Test musí být založen na výpočtu (inteligence jako výpočetní proces).
  - Definice inteligence musí být smysluplná, nikoliv „inteligence je to, co měří testy inteligence.“

C-Test

Hernandez-Orallo (2000):

- První přístup založený na *algoritmické teorii informace*.
- **Formální míra inteligence** zaměřená na faktor, který „nám umožňuje porozumět světu“.
- **C-Test** je **prakticky proveditelný test**.
- **Uvažuje pouze statické úlohy**.

Potřeba formální definice inteligence

- *Problematické vymezení inteligence*:
  - obtížně uchopitelný vysokoúrovňový koncept,
  - souvisí s dalšími obtížně uchopitelnými koncepty (učení, znalosti, komunikaci, jazyk, myšlení, kreativita),
  - na obecné úrovni shoda v psychologii, ale otevřené dílčí problémy,
  - IQ testy zřejmě postihují inteligenci (obecně × část, zaujatost?)
- *Způsobuje problémy disciplíně UI*:
  - odlišnosti strojů a lidí,
  - UI znamená pro každého něco jiného,
  - nedá se přiblížit k obecné UI a výzkum sklouzává k specifické UI,
- **Formální definice inteligence**:
  - zobecňující esenciální vlastnosti lidské inteligence,
  - umožňující jiné realizace,
  - založená na fundamentálních principech,
  - objektivní a převeditelná na praktický test.

Přirozená inteligence

- **Testy lidské inteligence**:
  - IQ testy měří něco důležitého (statisticky stálé, možnost predikce),
  - je to však inteligence, její typ, či jen nějaký aspekt?
  - otázka kulturní závislosti.
- **Testy inteligence u zvířat**:
  - abstrakce od lidského konceptu a metod testování,
  - porovnávání odlišných druhů,
  - odměňování namísto vysvětlení cíle testu.
- **Teorie inteligence v psychologii**:
  - řada více či méně oddělených schopností (Sternberg, Gardner),
  - obecná schopnost podkládající jiné (Spearman, Cattell).
- **Psychologické definice inteligence**:
  - vlastnost jedince interagujícího s prostředím, problémy a situacemi,
  - schopnost uspět (cíle, záměry, ...),
  - důraz na učení a přizpůsobení, tedy neznámost a novost prostředí.

Univerzální inteligence I

- Abstrahovaná **pracovní definice inteligence**:
  „Inteligence měří schopnost agenta vést si dobře v mnoha různých prostředích.“
- Formalizace:
  - **interakce agenta s prostředím**:
    - akce a_i,
    - vjemy (odměna r_i a pozorování o_i).
  - **agent** π:
    - funkce přiřazující historii interakcí nějakou akci,
    - podmíněná pravděpodobnostní míra nad prostorem akcí.
  - **prostředí** μ:
    - funkce přiřazující historii interakcí nějakou odměnu a pozorování,
    - vyčíslitelná podmíněná pravděpodobnostní míra nad prostorem odměn a pozorování.

Univerzální inteligence II

- Abstrahovaná **pracovní definice inteligence**:
  „Inteligence měří schopnost agenta vést si dobře v mnoha různých prostředích.“
- Formalizace (pokračování):
  - **míra úspěchu v prostředí**:
    - maximalizace očekávaných odměn,
    - suma odměn daných prostředím shora omezená 1,
    - časová preference řešení zabudovaná v prostředí.

$$V_\mu^\pi := \mathbb{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1$$

  - **prostor prostředí**:
    - vyčíslitelnost pravděpodobnostní míry prostředí umožňuje testy na počítači.

Univerzální inteligence III

- Abstrahovaná **pracovní definice inteligence**:
  „Inteligence měří schopnost agenta vést si dobře v mnoha různých prostředích.“
- Formalizace (pokračování):
  - **celková míra úspěchu**:
    - vliv jednotlivých prostředí podle jejich *algoritmické pravděpodobnosti* založené na složitosti (Kolmogorovská složitost K),
    - Occamova břitva – preference jednodušších hypotéz o prostředích před složitějšími.

$$2^{-K(\mu)}, \text{ kde} \qquad K(x) := \min_p\{l(p) : \mathcal{U}(p) = x\}$$

Univerzální inteligence IV

Legg – Hutter (2007):

- Abstrahovaná **pracovní definice inteligence**:
  „Inteligence měří schopnost agenta vést si dobře v mnoha různých prostředích.“
- Formální definice **univerzální inteligence**:

$$\Upsilon(\pi) := \sum_{\mu \in E} 2^{-K(\mu)} V_\mu^\pi \quad \text{kde} \qquad V_\mu^\pi := \mathbb{E}\left(\sum_{i=1}^{\infty} r_i\right) \leq 1$$

Univerzální inteligence různých agentů

- *náhodný* – náhodný výběr akce, neodhalí žádné pravidelnosti v prostředích, typicky nízká V_μ^π, nízká Υ,
- *specializovaný* – agent velmi dobrý v určité úloze (šachový program), mimo specializaci nízká V_μ^π, nízká Υ,
- *jednoduchý obecný* – statistika úspěšnosti akcí, náhodný průzkum, odhalí základní pravidelnosti prostředích, vyšší V_μ^π a Υ než předchozí,
- *obecný s historií* – korelace aktuální akce s historií interakcí, odhalí více pravidelností v prostředích, vyšší V_μ^π a Υ než předchozí,
- *obecný s historií a plánováním* – plánování dále než za aktuální odměnu, vyšší V_μ^π a Υ než předchozí,
- *velmi inteligentní agent* – dobré výsledky ve většině jednoduchých prostředí, ale i slušné výsledky v mnoha komplexních prostředích, většinou vysoká V_μ^π,
- *superinteligence* – vždy vybere akci s nejvyšší budoucí odměnou, dokonalá predikce, maximální V_μ^π,
- *člověk* – nakolik je omezen evoluční adaptací na určitá prostředí?

Optimální agent AIXI

Legg – Hutter (2007) a Hutter (2012) – rovnice AIXI:

$$a_k := \arg\max_{a_k} \sum_{o_k r_k} \ldots \max_{a_m} \sum_{o_m r_m} [r_k + \ldots + r_m] \sum_{q:\ U(q,a_1 \ldots a_m) = o_1 r_1 \ldots o_m r_m} 2^{-l(q)}$$

„Sequential decision theory (Bellman's equation) formally solves the problem of rational agents in uncertain worlds if the true environmental probability distribution is known. If the environment is unknown, Bayesians replace the true distribution by a weighted mixture of distributions from some (hypothesis) class. Using the large class of all (semi)measures that are (semi)computable on a Turing machine bears in mind Epicurus, who teaches not to discard any (consistent) hypothesis. In order not to ignore Ockham, who would select the simplest hypothesis, Solomonoff defined a universal prior that assigns high/low prior weight to simple/complex environments, where Kolmogorov quantifies complexity."

Vlastnosti Univerzální inteligence

- **validní** – definuje inteligenci, ne nějakou související kvalitu, nebo jen aspekt inteligence,
- **smysluplná** – absolutní porovnatelné měřítko (skalární hodnota),
- **široký rozsah** – od velmi nízké po velmi vysokou inteligenci,
- **obecná** – srovnání značně odlišných (nejlépe všech) agentů,
- **nepředpojatá** – testové úlohy nejsou výsledkem kulturní nebo jiné předpojatosti, ale závisí na referenčním Turingově stroji $\mathcal{U}$,
- **fundamentální** – založena na „pevných základech“ (koncepty výpočtu, informace a složitosti),
- **formální** – jasně definovaná matematická rovnice,
- **objektivní** – nezávisí na subjektivních kritériích,
- **univerzální** – není antropocentrická,
- **není praktickým testem** – Kolmogorovská složitost je nevyčíslitelná, převod na test bude vyžadovat aproximaci.

Námitky k Univerzální inteligenci I

Je to zjevně špatně

- *Definovat inteligenci vzorcem prostě nejde.*
- Legg a Hutter:
  - ...

Je to zjevně správně, nic nového

- *Definice je vlastně popis „reinforcement learning“.*
- Legg a Hutter:
  - Spojení „reinforcement learning“, Occamovy břitvy, teorie složitosti.
  - Navázání na neformální definici inteligence, prokázání, že má správné vlastnosti, propojení s teorií univerzálně optimálních učících agentů, ...

Námitky k Univerzální inteligenci II

Předpoklad vyčíslitelnosti prostředí příliš silný

- *Vesmír není vyčíslitelný.*
- Legg a Hutter:
  - Vyčíslitelnost ve smyslu vyčíslitelné pravděpodobnostní míry nad budoucími událostmi.
  - Nejsou důkazy pro nevyčíslitelnost ve fyzice v tomto smyslu.
  - Vše dosud známé je v tomto smyslu vyčíslitelné.

Prostředí vracející odměnu s omezeným součtem je nerealistický

- *Vesmír nevrací odměny.*
- Legg a Hutter:
  - Jde o testovací framework, subjektivní odměny by neumožnily testovat.

Kritika Univerzální inteligence

Hibbard (2009):

- Kolmogorovská složitost K činí míru závislou na volbě *referenčního Turingova stroje* $\mathcal{U}$.
- Definice je **předpojatá vůči malé množině prostředí** popsaných krátkými programy.
- Míře inteligence **dominují výsledky v těchto prostředích**.
- Pro různé $\mathcal{U}$ mohou být dominující prostředí značně odlišná.

Pragmatická obecná inteligence

Goertzel (2010):

- *Kritika univerzální inteligence*:
  - reální **agenti** jsou **přizpůsobení některým prostředím** – „biased generality“ namísto univerzálnosti,
  - ne všechno inteligentní chování je o vyhledávání odměny,
  - **odměny a cíle** jsou často **dané samotným agentem**,
  - externí odměny jako testovací framework pro dedukci inteligence.
- **Pragmatická obecná inteligence:**
  - lze zvolit pravděpodobnostní distribuci prostředí,
  - podmíněná distribuce cílů v prostředích,
  - vyhodnocování po časový interval od volby cíle do jeho dosažení.

„Inteligence měří schopnost agenta dosahovat komplexních cílů v komplexních prostředích.“

$$\Pi(\pi) \equiv \sum_{\mu \in E,\, g \in G,\, T} \nu(\mu)\gamma(g,\,\mu)\, V^{\pi}_{\mu,\,g,\,T}, \quad \text{kde} \quad V^{\pi}_{\mu,\,g,\,T} \equiv \mathbb{E}\left(\sum_{i=s}^{t} r_g(I_{g,s,i})\right)$$

Efektivní pragmatická obecná inteligence

Goertzel (2010):

- *Kritika univerzální inteligence*:
  - reální **agenti** operují **s omezenými zdroji**.
- **Efektivní pragmatická obecná inteligence**:
  - normalizace pragmatické obecné inteligence výpočetními nároky agenta,
  - podmíněná distribuce spotřeby výpočetních zdrojů.

$$\Pi_{\mathrm{Eff}}(\pi) \equiv \sum_{\mu \in E,\, g \in G,\, Q,T} \frac{\nu(\mu)\gamma(g,\ \mu)\eta_{\pi,\mu,g,T}(Q)}{Q} V^{\pi}_{\mu,g,T}$$

Intelektuální šíře agenta

Goertzel (2010):

- *Kritika univerzální inteligence*:
  - **chybí kvantifikace obecnosti inteligence**.
- **Intelektuální šíře agenta**:
  - fuzzy množina agentů, kteří jsou inteligentní vůči kontextu (cíl, prostředí, časový interval),
  - fuzzy množina kontextů, vůči kterým jsou agenti inteligentní,
  - normalizace fuzzy množiny na pravděpodobnostní distribuci,
  - entropiie pravděpodobnostní distribuce jako míra intelektuální šíře.
- **specifická UI** – agent s nízkou intelektuální šíří,
- **obecná UI – agent** s vysokou intelektuální šíří.

Kdykoliv přerušitelný test inteligence

Hernández-Orallo – Dowe (2010):

- *Návrh* testu inteligence pro:
  - současné i budoucí umělé i biologické agenty,
  - různé úrovně inteligence i různé časové škály.
- *Dosažení vyčíslitelnosti Univerzální inteligence:*
  - konečný vzorek prostředí – problém **diskriminační síly prostředí** vyřešen požadavkem **citlivosti vůči odměně**,
  - konečný počet interakcí mezi agentem a prostředím – **mění způsob výpočtu skóre na průměrování** a **požaduje vyvážená prostředí**,
  - složitostní funkce inspirovaná Levinovou *Kt* zachová Occamovu břitvu.
- **Fyzický čas** a **adaptivní testovací procedura**.

Insa-Cabrera et al. (2011):

- (příliš) zjednodušený *prototyp* a *demo experimenty*,
- **druhově specifická rozhraní** umožňují použít stejný test.

Test algoritmického IQ I

Legg – Veness (2013):

- aproximace *univerzální inteligence* – **test algoritmického IQ**:

$$\hat{\Upsilon}(\pi) := \frac{1}{N}\sum_{i=1}^{N}\hat{V}_{p_i}^{\pi}, \text{ kde } \hat{V}_{p_i}^{\pi} := \frac{1}{k}\sum_{i=1}^{k} r_i \text{ a } p_i \text{ dle } M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}$$

  - **Konečný vzorek programů prostředí**
  - **Konečný počet interakcí** mezi agentem a prostředím.
  - Solomonoffova **univerzální distribuce** jako alternativní vyjádření Occamovy břitvy (preference kratších programů).
  - **Omezená výše odměny za jednu interakci** [-100;+100], ale už ne omezení součtu odměn.

Test algoritmického IQ II

Legg – Veness (2013):

- Programy prostředí:
  - **Počítají aktuální odměnu a pozorování** z historie interakcí.
  - Jsou **náhodně generované Turingovsky úplné programy**
  - Jednoduchý referenční Turingův stroj: **rozšířený BF stroj**.
  - Pouze programy prostředí, které skončí v rozumném čase.
  - **Jsou vyvážené**, náhodně se chovající agent získá skóre 0.
  - **Některé programy bez diskriminační síly jsou vynechány** ze vzorku.

Test algoritmického IQ III

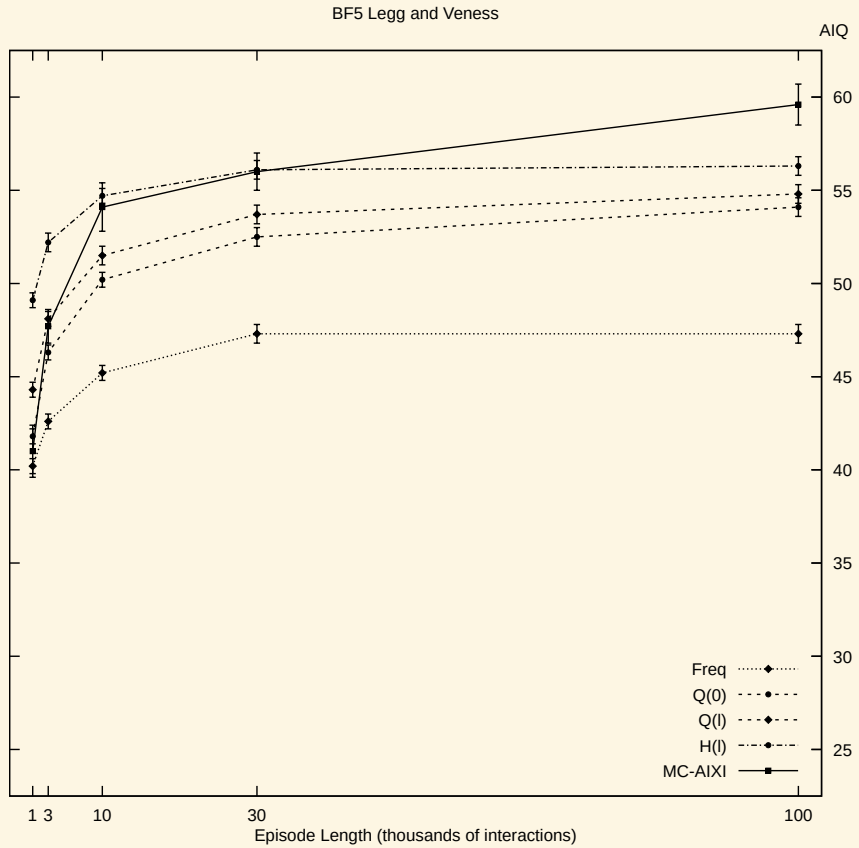
Legg – Veness (2013); Müller (1993):

- **Rozšířený BF referenční stroj**:
  - `+-` zvýší/sníží symbol v pracovní pásce,
  - `,.` přečte symbol ze vstupní pásky/zapíše symbol na výstupní pásku,
  - `<>` posune pracovní pásku o políčko doleva/doprava,
  - `[]` pokud je symbol na pracovní pásce nenulový začne cyklus/skončí cyklus,
  - `%` zapíše náhodný symbol na pracovní pásku,
  - `#` skončí program.

```
[+.>]-,,#
+[[[+>]<<.,<%->]%,>-<],+[+>.,.%]+>.%%....<+#
```

Test algoritmického IQ IV

- *Open source implementace:*
  - odhad pro 200 000 programů prostředí z 10 000 vzorků (pro MC-AIXI 1 000 vzorků),
  - test po 1, 3, 10, 30 a 100 tisících interakcích,
  - výchozí nastavení BF stroje.
- *Agenti v testu:*
  - *random* – random behavior,
  - *freq* – chooses action with highest average reward,
  - Q_0 – basic Q-learning,
  - $Q\lambda$ – Q-learning with eligibility traces,
  - $HLQ\lambda$ – Q-learning with automatic learning rate,
  - *MC-AIXI* – wrapper for Monte Carlo approximation of AIXI.



Obrázek: AIQ různých agentů

Míra obtížnosti prostředí

Hernández-Orallo (2015, 2017):

- **Obtížnost problému či prostředí** závisí na **složitosti řešení** namísto složitosti popisu prostředí.
- Obtížnost prostředí je určena **obtížností nejjednodušší řešení** podle nějaké obtížnostní funkce.
- *Obtížnostní funkce založená na Levinově složitosti* uvažuje *délku řešení* a také *počet výpočetních kroků.*
- *Změna způsobu zprůměrování výsledků agentů:*
  1. uniformní nebo klesající distribuce obtížností,
  2. uniformní distribuce řešení o dané obtížnosti,
  3. univerzální distribuce úloh s daným řešením.
- **Závislost** na volbě referenčního stroje **je snížena**.

Teorie úloh

Thórisson et al. (2015, 2016):

- **Úlohy či prostředí jsou klíčové pro vyhodnocení inteligence**, jejich formalizace je však omezená (interakce s agentem, složitost, obtížnost).
- **Potřeba důkladně rozumět úlohám** zejména, jak je porovnat, variovat a vytvářet.
- *Seznam požadavků na teorii úloh*, aby bylo možné **modelovat fyzické úlohy** s časovými a energetickými omezeními.
- *Počáteční formalizmus pro analýzu úloh* (proměnné, domény, invariantní relace, dynamické funkce).

Diskuze

- Skupiny přístupů vyhodnocující inteligenci s odlišnými předpoklady:
  1. **Úspěch ve složité úloze je postačující podmínkou pro inteligenci** (Turingův test) *postačitelnost lze zpochybnit pro nelidské subjekty.*
  2. **Explicitní ověření úspěchu v jednoduchých i složitých prostředích je nutné pro přiznání inteligence** (AIQ test) *dostatečně robustní pro lidi, umělé systémy a zvířata.*

Závěr

- Souvislosti **myšlení, jazyka a rozumění**.
  - **Turingův test** a jeho rozšířené verze,
  - **čínský pokoj**,
  - **kognitivní architektury**.
- Formální přístupy založené na **algoritmické teorii informace**:
  - Definice **univerzální inteligence**,
  - *Kdykoliv přerušitelný test inteligence* a **AIQ test**.
- Přehledový článek v češtině: Vadinský (2018).

Zdroje I

Anderson, J. R. et al. An Integrated Theory of the Mind. *Psychological Review*. 2004, vol. 111, no. 4, s. 1036--1060. ISSN 0033-295X. doi: 10.1037/0033-295x.111.4.1036.

Besold, T. R. – Hernández-Orallo, J. – Schmid, U. Can Machine Intelligence be Measured in the Same Way as Human intelligence? *KI – Künstliche Intelligenz*. 2015, vol. 29, no. 3, s. 291--297. doi: 10.1007/s13218-015-0361-4.

Bringsjord, S. – Schimanski, B. What Is Artificial Intelligence? Psychometric AI as an Answer. In Gottlob, G. (Ed.) *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI'03)*, s. 887--893, 2003.

Zdroje II

Burge, T. Individualism and the Mental. *Midwest studies in philosophy*. 1979, vol. 4, no. 1, s. 73--121. ISSN 1475-4975. doi: 10.1111/j.1475-4975.1979.tb00374.x.

Mey, M. *The Cognitive Paradigm*. Chicago and London: University of Chicago Press, 1992. doi: 10.1007/978-94-009-7956-7. ISBN 0-226-14259-0.

Descartes, R. *Rozprava o metodě*. Praha: Svoboda, 3. vyd., 1637. 1992 (poprvé publikováno 1637).

Goertzel, B. Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*. 2014, vol. 5, no. 1, s. 1--48. doi: 10.2478/jagi-2014-0001.

Zdroje III

Goertzel, B. Toward a Formal Characterization of Real-World General Intelligence. In Baum, E. – Hutter, M. – Kitzelmann, E. (Ed.) *Proceedings of the 3rd International Conference on Artificial General Intelligence (AGI 2010), Lugano, Switzerland*, 11 / *Advances in Intelligent Systems Research*, s. 19--24, Amsterdam-Beijing-Paris, 2010. Atlantis Press. doi: 10.2991/agi.2010.17. ISBN 978-90-78677-36-9.

Harnad, S. Other Bodies, Other Minds: A Machine Incarnation of an old Philosophical Problem. *Minds and Machines*. 1991, vol. 1, no. 1, s. 43--54. ISSN 0924-6495. doi: 10.1007/BF00360578.

Zdroje IV

Havel, I. M. Přirozené a umělé myšlení jako filozofický problém. In Mařík, V. – Štěpánková, O. – Lažanský, J. (Ed.) *Umělá inteligence 3*. Praha: Akademia, 1. vyd., 2001. s. 17--75. ISBN 80-200-0472-6.

Hernández-Orallo, J. *The Measure of All Minds*. Cambridge: Cambridge University Press, 1. vyd., 2017. doi: 10.1017/9781316594179. ISBN 978-1-10715-301-1.

Hernandez-Orallo, J. Beyond the Turing Test. *Journal of Logic, Language and Information*. 2000, vol. 9, no. 4, s. 447--466. ISSN 0925-8531. doi: 10.1023/A:1008367325700.

Hernández-Orallo, J. On environment difficulty and discriminating power. *Autonomous Agents and Multi-Agent Systems*. 2015, vol. 29, no. 3, s. 402--454. ISSN 1387-2532. doi: 10.1007/s10458-014-9257-1.

Zdroje V

Hernández-Orallo, J. – Dowe, D. L. Measuring Universal Intelligence: Towards an Anytime Intelligence Test. *Artificial Intelligence*. 2010, vol. 174, no. 18, s. 1508--1539. ISSN 0004-3702. doi: 10.1016/j.artint.2010.09.006.

Hibbard, B. Bias and No Free Lunch in Formal Measures of Intelligence. *Journal of Artificial General Intelligence*. 2009, vol. 1, no. 1, s. 54--61. doi: 10.2478/v10229-011-0004-6.

Hutter, M. One Decade of Universal Artificial Intelligence. In Wang, P. – Goertzel, B. (Ed.) *Theoretical Foundations of Artificial General Intelligence*, 4 / *Atlantis Thinking Machines*. Paris: Atlantis Press, 2012. s. 67--88. doi: 10.2991/978-94-91216-62-6_5. ISBN 978-94-6239055-3.

Zdroje VI

Hyslop, A. Other Minds. In Zalta, E. N. (Ed.) *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2014. vyd., 2014. Dostupné z: <http://plato.stanford.edu/archives/spr2014/entries/other-minds/>.

Insa-Cabrera, J. et al. Comparing Humans and AI Agents. In Schmidhuber, J. – Thórisson, K. R. – Looks, M. (Ed.) *Proceedings of the 4th International Conference on Artificial General Intelligence (AGI 2011), Mountain View, USA*, 6830 / *Lecture Notes in Artificial Intelligence*, s. 122--132, Berlin, 2011. Springer. doi: 10.1007/978-3-642-22887-2_13. ISBN 978-3-642-22887-2.

Zdroje VII

Kripke, S. A. *Naming and necessity*. Cambridge: Harvard University Press, 1972. ISBN 978-0-674-59846-1.

Legg, S. – Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*. Dec 2007, vol. 17, no. 4, s. 391--444. ISSN 0924-6495. doi: 10.1007/s11023-007-9079-x.

Legg, S. – Veness, J. An Approximation of the Universal Intelligence Measure. In Dowe, D. L. (Ed.) *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, 7070 / *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2013. s. 236--249. doi: 10.1007/978-3-642-44958-1_18.

Zdroje VIII

Müller, U. *dev/lang/brainfuck-2.lha in Aminet* [online]. 9. 6. 1993. [cit. 2017-06-26]. Dostupné z: <http://aminet.net/package.php?package=dev/lang/brainfuck-2.lha>.

Pstružina, K. Turing v Searleově čínské místnosti. *Acta oeconomica Pragensia*. 2003, vol. 11, no. 8, s. 9--16.

Putnam, H. *The Meaning of 'Meaning'*, s. 215--271. Cambridge: Cambridge University Press, 1975. ISBN 978-0-521-20668-6.

Schweizer, P. The Externalist Foundations of a Truly Total Turing Test. *Minds and Machines*. Aug 2012, vol. 22, no. 3, s. 191--212. ISSN 0924-6495. doi: 10.1007/s11023-012-9272-4. http://www.research.ed.ac.uk/portal/en/persons/paul--schweizer%2894664317-391f-4731-9125-577d61b109ba%29.html.

Zdroje IX

Searle, J. R. Minds, Brains, and Programs. *Behavioral and Brain Sciences*. 1980, no. 3, s. 417--457. doi: 10.1017/S0140525X00005756.

Sun, R. The Importance of Cognitive Architectures: An Analysis Based on CLARION. *Journal of Experimental & Theoretical Artificial Intelligence*. 2007, vol. 19, no. 2, s. 159--193. ISSN 0952-813X. doi: 10.1080/09528130701191560.

Thomsen, K. The Cerebellum in the Ouroboros Model, the "Interpolator Hypothesis". In Shimizu, S. – Bossomaier, T. (Ed.) *Proceedings of the 5th International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2013), Valencia, Spain*, s. 37--41, Wilmington, 2013. IARIA. Dostupné z: <http://www.thinkmind.org/download.php?

Zdroje X

articleid=cognitive_2013_2_30_40069>. ISBN 978-1-61208-273-8.

Thórisson, K. R. et al. Towards Flexible Task Environments for Comprehensive Evaluation of Artificial Intelligent Systems and Automatic Learners. In Bieger, J. – Goertzel, B. – Potapov, A. (Ed.) *Proceedings of the 8th International Conference on Artificial General Intelligence (AGI 2015), Berlin, Germany*, 9205 / *Lecture Notes in Artificial Intelligence*, s. 187--196, Berlin, 2015. Springer. doi: 10.1007/978-3-319-21365-1_20. ISBN 978-3-319-21364-4.

Zdroje XI

Thórisson, K. R. et al. Why Artificial Intelligence Needs a Task Theory And What It Might Look Like. In Steunebrink, B. – Wang, P. – Goertzel, B. (Ed.) *Proceedings of the 9th International Conference on Artificial General Intelligence (AGI 2016), New York, USA*, 9782 / *Lecture Notes in Artificial Intelligence*, s. 118--128, New York, 2016. Springer. doi: 10.1007/978-3-319-41649-612. ISBN 978-3-319-41649-6.

Turing, A. M. Computing Machinery and Intelligence. *Mind*. 1950, vol. 59, no. 236, s. 433--460. ISSN 0026-4423.

Vadinský, O. Přehled přístupů k vyhodnocování inteligence umělých systémů. *Acta Informatica Pragensia*. 2018, vol. 7, no. 1, s. 74--103. ISSN 1805-4951. doi: 10.18267/j.aip.115.

Zdroje XII

Wang, P. On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*. 2019, vol. 2, no. 10, s. 1--37. doi: 10.2478/jagi-2019-0002.