Acta informatica Pragensia, 2018, 7(1): 74–103
DOI: 10.18267/j.aip.115
Original Article

Přehled přístupů k vyhodnocování inteligence umělých systémů

An Overview of Approaches Evaluating Intelligence of Artificial Systems

Ondřej Vadinský [1]

Abstrakt

Obecná umělá inteligence usiluje o vytvoření umělých systémů schopných řešit mnoho různých, a to i během vývoje nepředvídaných, úloh, což takové systémy činí svou inteligencí srovnatelné s lidmi. To však vyžaduje existenci vhodných metod vyhodnocujících, zda a nakolik jsou umělé systémy inteligentní. Tento přehledový článek hledá právě takové evaluační metody. Provádí proto rozsáhlou rešerši literatury pokrývající jak filosofické a kognitivní předpoklady inteligence, tak i formální definice a praktické testy vycházející z algoritmické teorie informace. Na základě porovnání představených metod článek odhaluje dvě rozdílné skupiny přístupů založené na principiálně odlišných předpokladech. Zatímco starší přístupy, jako např. Turingův test, jsou založeny na předpokladu, že úspěch v komplexní činnosti je postačující pro přiznání inteligence, nové přístupy, jako např. test algoritmického IQ, kromě toho vyžadují i důkladné ověření úspěšnosti v jednoduchých činnostech. V důsledku tohoto zjištění článek dochází k závěru, že test algoritmického IQ založený na definici univerzální inteligence je v současné době nejlepším kandidátem na vhodný prakticky proveditelný test obecné inteligence umělých systémů. Ačkoliv i tento test má několik známých limitů.

Klíčová slova: Obecná umělá inteligence, definice univerzální inteligence, kdykoliv přerušitelný test inteligence, test algoritmického IQ, vyhodnocování inteligence umělých systémů.

Abstract

Artificial General Intelligence seeks to create an artificial system capable of solving many different and possibly unforeseen tasks thus being comparable in its intelligence to that of a human. Such an endeavour, however, requires suitable methods that can evaluate whether an artificial system is intelligent, and to what extent. This review paper searches for such evaluation methods. Therefore, an extensive literature overview is conducted that covers both philosophical and cognitive presumptions of intelligence as well as formal definitions and practical tests of intelligence grounded in Algorithmic Information Theory. Based on a comparison of the introduced approaches, the paper identifies two distinct groups based on fundamentally different presumptions. The one group of approaches, such as Turing test, is based on the presumption that success in a complex task is a sufficient condition for intelligence evaluation, while the other group of approaches, such as Algorithmic Intelligence Quotient test, also require explicit verification of success in simple tasks. This paper, therefore, concludes that the Algorithmic Intelligence Quotient test, derived from Universal Intelligence definition, is currently the most suitable candidate for a practical intelligence evaluation method of artificial systems. Although the test has several known limitations.

Keywords: Artificial General Intelligence, Universal Intelligence Definition, Anytime Intelligence Test, Algorithmic Intelligence Quotient Test, Evaluating Intelligence of Artificial Systems.

[1]Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
✉ ondrej.vadinsky@vse.cz

1 Úvod

Otázka, jak poznat a vyhodnotit, zda je umělý systém inteligentní, stála již u zrodu umělé inteligence jako vědecké disciplíny (Turing, 1950). Tato otázka úzce souvisí s definicí a chápáním inteligence či myšlení a jde tak o multidisciplinární, nikoliv úzce informatický problém. Tento původní směr ve vývoji umělé inteligence však byl postupně upozaděn zaměřením se na řešení specifických problémů a dílčích úloh. Jak argumentují například Legg & Hutter (2007b), je pro vývoj v oboru umělé inteligence dostatečně přesná definice klíčová tím, že umožňuje tento vývoj směřovat, a nejde tak o pouhé tápání ve tmě, byť i to dokáže přinést zajímavé výsledky. Až v nedávné době se snaha inteligenci přesněji definovat a prakticky vyhodnocovat opět stává předmětem podrobnějšího výzkumu na poli umělé inteligence. Přispívá k tomu zejména rozvoj dílčího oboru označovaného jako *obecná umělá inteligence*, jehož cílem je vývoj uměle inteligentních systémů, které budou rozsahem svých schopností a svou univerzálností srovnatelné s lidskou inteligencí (Goertzel, 2014). Postupně tak krystalizuje nová výzkumná oblast zaměřená na *univerzální vyhodnocování inteligence* (Hernández-Orallo, 2017).

V tomto přehledovém článku si kladu dva cíle: 1) Představit jednotlivé přístupy k vyhodnocování inteligence umělých systémů a 2) na základě jejich porovnání vybrat vhodný přístup pro další empirický výzkum. Ač článek sleduje téma vyhodnocování inteligence umělých systémů od jeho počátku, největší pozornost je věnována období po roce 2000, kdy postupně nastává renesance zájmu o toto téma. Zejména tento nedávný vývoj pak dosud není v českém jazyce obšírněji zpracován.

Otázka po vyhodnocování inteligence umělých systémů má široké souvislosti přesahující pro obor umělé inteligence primární otázky týkající se porovnání konkrétních systémů, či jejich různých vylepšení. Jde totiž i o vyjasnění toho, co je primárním cílem disciplíny umělé inteligence, co tento cíl vlastně znamená, jak se k němu dostat a nakolik se mu ten který systém přibližuje. Odpovědi na tyto otázky mohou značnou měrou přispět k průběhu v současné době se pozvolna otevírající debaty ohledně vlivu a dopadů zavádění uměle inteligentních systémů na společnost. Už jen vyjasnění si toho, zda umělá inteligence směřuje k tvorbě „chytrých nástrojů“, či „samostatně uvažujících svobodně jednajících entit“ a kam které systémy spadají, resp. nakolik tohoto cíle dosahují, může vnést nové perspektivy do diskuze o etických aspektech, podobě ekonomické transformace, či proměně společenských vztahů způsobené takto pojatou umělou inteligencí. Zároveň, ač se v článku zabývám vyhodnocováním inteligence umělých systémů, dotýká se toto téma pro člověka natolik niterné záležitosti jako je myšlení respektive inteligence. Nejde tedy jen o samotnou konstrukci inteligentních strojů, ale i o jakési nastavení zrcadla člověku.

Primární metodou použitou v tomto článku je rešerše literatury dostupné ke zkoumané problematice. Ta vzhledem k obecnosti konceptů inteligence a myšlení pokrývá široké spektrum oborů kognitivní vědy a není jen úzce informatická. Téma je také zastoupeno nejen v nedávné literatuře ale i ve starších zdrojích z doby kolem vzniku disciplíny umělé inteligence. Rešerši jsem provedl v databázích Web of Science, Scopus, ProQuest Central, JSTOR a arXiv, přičemž jsem využil zejména následující klíčová slova (v angličtině, zde uvádím i české překlady): obecná umělá inteligence (artificial general intelligence), měření inteligence (measurement of intelligence), vyhodnocování umělé inteligence (artificial intelligence evaluation), Turingův test (Turing test) a univerzální inteligence (Universal Intelligence). Pro posouzení vhodnosti představených přístupů a výběr konkrétního přístupu pro další empirický výzkum pak článek využije analýzu a syntézu poznatků získaných z této rešerše.

Skrze filosofickou reflexi a znalosti kognitivních věd lze provést uchopení pojmu inteligence

a jeho usouvztažnění s dalšími vysokoúrovňovými pojmy popisující související schopnosti mysli, jak ukáže sekce 2. Ač takový přístup poskytuje prvotní náhled na řešenou problematiku, nepřináší však podrobně specifikované formální definice a často ani prakticky realizovatelné testy. Nedávnou snahu dospět k formální definici inteligence ukotvené v *algoritmické teorii informace* bude sledovat sekce 3. Na základě formálních definic pak lze budovat prakticky proveditelné testy, čemuž se bude věnovat sekce 4. Posouzení představených přístupů k vyhodnocování inteligence umělých systémů provede sekce 5.

2 Filosofické a kognitivní předpoklady inteligence

Tato sekce provede prvotní náhled na pojem inteligence založený na reflexi tohoto pojmu u vybraných autorů ve filosofii, v umělé inteligenci a v dalších kognitivních vědách. Jako východisko poslouží perspektiva umělé inteligence, tedy snahy o vytvoření umělého inteligentního systému. Odhalené poznatky jsou z deskriptivního pohledu vlastnostmi inteligence, z pohledu konstrukce umělých systémů však jde o předpoklady, které systém musí splňovat, aby byl inteligentní.

Nejprve sekce 2.1 vymezí různá chápání hlavního cíle disciplíny umělé inteligence. Na linii *Turingova testu*, jeho východisek a snah o jeho rozšíření sledované v sekci 2.2 budou ukázána různá filosofická uchopení pojmu inteligence. Vztahu inteligence a dalších kognitivních schopností se bude věnovat sekce 2.3.

2.1 Vymezení disciplíny umělá inteligence

Hlavní cíl disciplíny umělá inteligence (UI) lze chápat různě. Z filosofie přichází rozlišení mezi *silnou* a *slabou umělou inteligencí*, které uvede sekce 2.1.1, v samotné komunitě umělé inteligence se poněkud později objevilo nyní preferované rozlišení na *obecnou* a *specifickou umělou inteligenci*, které popíše sekce 2.1.2.

2.1.1 Vymezení silná – slabá umělá inteligence

Searle (1980) rozlišuje umělou inteligenci podle toho, zda si za cíl klade vysvětlit a duplikovat mysl (*silná UI*), nebo jde pouze o nástroj pro modelování a simulování mysli (*slabá UI*). V případě *silné UI* tak má počítač vykonávající správný program kognitivní stavy (zejména rozumění), zatímco v případě *slabé UI* jde jen o přesněji formulovanou hypotézu o mysli. *Silnou UI* pak Searle identifikuje zejména s tradičním symbolickým přístupem k umělé inteligenci, který je tak chápán jako vysvětlení fungování mysli.

Tuto distinkci však lze interpretovat poněkud volněji, pokud se zaměříme na její normativní aspekt namísto aspektu deskriptivního. Za *silnou UI* lze považovat takový systém, který by byl srovnatelně mocný jako lidská inteligence – zejména co do její univerzálnosti a přítomnosti dalších kvalit, které s inteligencí u lidí spojujeme, např. rozumění. Oproti tomu *slabá UI* je systém, který by srovnatelné univerzálnosti nedosahoval, a šlo by pouze o užitečný nástroj pro řešení konkrétních problémů.

2.1.2 Vymezení obecná – specifická umělá inteligence

Výše zmíněná interpretace Searla je blízká poměrně nedávnému rozlišení *specifická – obecná umělá inteligence*, jak jej vymezuje např. Goertzel (2014). *Specifická UI* nabízí systémy implementující konkrétní schopnostmi odtržené od dalších, které s nimi souvisejí, či umožňující řešit jednu či několik úzce vymezených úloh, které mohou být jednoduché, ale také značně sofistikované. Mimo tuto svou jasně vymezenou množinu úloh však *specifická UI* bez často rozsáhlého zásahu svého tvůrce selže.

Obecná UI naopak usiluje o vytvoření systémů, které jsou schopné obecného řešení problémů, resp. obecného inteligentního jednání. Na základě toho pak dokáží řešit širokou množinu úkolů ať již jednoduchých, nebo složitých, a to v mnoha různých situacích a prostředích. S *obecnou umělou inteligencí* bývá také spojovaná schopnost přenášet znalosti z řešení jedné úlohy do řešení jiné úlohy, se kterou se systém dosud nesetkal, či k jejíž řešení nebyl původně určen. Prakticky to však vzhledem k omezeným zdrojům systému neznamená neomezenou obecnost, ale různou míru omezené obecnosti, při které lze rozlišit úlohy, v jejichž řešení je systém efektivnější než v řešení jiných (Goertzel, 2014).

Za typický příklad *specifické UI* lze označit šachové programy (Levy & Newborn, 1991). Již od vítězství stroje *Deep Blue* nad Kasparovem jde o systémy překonávající ve své specifické doméně i ty nejlepší z lidí. Postavíme-li však šachový program před jakoukoliv jinou úlohu, ať už podobnou jako hru v piškvorky, či značně odlišnou jako ovládání robotického vozítka na Marsu, selžou. Oproti tomu *obecná UI* se tento nový problém naučí řešit, podobně jako se jej dokáže naučit i člověk. Příkladem technologie mající rysy *obecné UI* jsou výstupy výzkumu společnosti *Deep Mind*: agent založený na hlubinných *Q*-sítích schopný naučit se efektivně hrát řadu různých her pro Atari (Mnih et al., 2015), či systém *Alpha Zero*, který hrou sám se sebou dokázal bez předchozích znalostí dosáhnout mistrovské úrovně v šachu, šógi a go (Silver et al., 2017).

V tomto článku budu nadále nahlížet na umělou inteligenci z perspektivy *obecné UI*. Uváděné metody vyhodnocující umělou inteligenci sice lze vztáhnout i na *specifické umělé inteligence*, zde se však plně projeví jejich nedostatečnost.

2.2 Turingův test a jeho rozšíření

Pro umělou inteligenci je klíčovou otázkou: „Co je to myšlení?“ Její uchopení lze najít už u Descarta, jak ukáže sekce 2.2.1. Známějším je však Turingovo zpracování, pro něž se ujalo označení *Turingův test*, kterým se bude zabývat sekce 2.2.2. Ač se obor umělá inteligence ubíral spíše jiným směrem, lze najít přístupy, které Turingovu myšlenku dále rozpracovávají. Jde zejména o *Úplný Turingův test* přiblížený v sekci 2.2.3 a o *Skutečně úplný Turingův test* uvedený v sekci 2.2.4.

2.2.1 Descartovo rozlišení stroje a člověka

Descartes (1637) věnuje jednu pasáž své *Rozpravy o metodě* rozlišení stroje a člověka. Všímá si zde dvou vlastností, které podle něj náleží pouze člověku: univerzálnost myšlení a schopnost rozumné řeči.

> *„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. 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." (Descartes, 1637, str. 41)

Pro další výklad je zejména důležité Descartovo uchopení univerzálnosti myšlení jako všestrannosti rozumu, tedy schopnosti řešit různé problémy, byť ne nutně nejlépe. Zde lze spatřovat paralely s výše uváděnými definicemi *obecné*, resp. *silné umělé inteligence*. Druhým zásadním momentem je pak spojení inteligence s rozumnou řečí, kterou Descartes (1637) chápe jako schopnost odpovídat na všemožné otázky. Právě tento aspekt uvedené pasáže předznamenává mnohem pozdější Turingovo dílo.

2.2.2 Turingův test

Turing (1950) si klade otázku: „Mohou stroje myslet?" Všímá si však nejasnosti v ní obsažených pojmů, zejména pak pojmu myšlení. Provádí proto následující obrat: Ptá se, zda stroje [1] mohou napodobit lidské chování a myšlení do takové míry, kdy běžný člověk nedokáže rozlišit, zda rozmlouvá s člověkem, či se strojem. Pro toto pojetí se později vžil název *Turingův test*. Descartem uvedené vlastnosti jsou v něm implicitně přítomny, schopnost rozumné řeči je pak pro úspěch stroje v testu Turingem považována za klíčovou.

Turing (1950) založil svůj test na imitační hře, která předpokládá fyzické oddělení účastníků, tedy dvou hráčů: člověka a stroje, a jednoho rozhodčího: člověka. Cílem rozhodčího je určit, kdo je člověk a kdo stroj na základě nepřímé komunikace s hráči přes terminál. Jediné, co tedy může rozhodčí při svém rozhodování využít, jsou odpovědi na jeho otázky. Obsah otázek však není omezen a lze tak jejich prostřednictvím zkoumat i jiné než jen jazykové schopnosti stroje. Zde se tedy může projevit ona zmiňovaná univerzálnost myšlení.

Díky požadavku na nepřímou komunikaci mezi účastníky testu nemusí stroj napodobovat člověka svým vzhledem, jak uvažuje Descartes (1637). Turing (1950) tak ve svém testu nerozlišuje účastníky na základě povrchních znaků, ale až na základě případných odlišností v projevech inteligence (myšlení). Pokud však rozhodčí nedokáže rozlišit, zda komunikuje s člověkem, či se strojem, vyplývá z toho, podle Turinga, že stroj je s člověkem srovnatelný, ba dokonce že je lepší než člověk, protože překonává neúspěšného hráče, a tak jsme povinni i stroji připsat lidem běžně připisované atributy jako mysl či vědomí.

Turing (1950) také skrze ideu učících se strojů předkládá návrh, jak dosáhnout úspěchu v testu. Absenci správného programu pro složitou „dospělou mysl" tak chce překonat naprogramováním jednodušší „dětské mysli" a jejím vystavením světu a společnosti. Pro „dětskou mysl" jsou důležité zejména procesy učení, které se při jejím vystavení světu uplatní a umožní jí, aby dosáhla úrovně „dospělé mysli". Turing tímto upozorňuje na fakt, že řada lidských schopností je naučená, a také na souvislost myšlení a inteligence s učením.

Myšlenky obsažené v Turingově článku se staly jedním z východisek funkcionalizmu, jednoho ze směrů filosofie mysli, který za určující vlastnost mentálních stavů nepovažuje jejich vnitřní konstituci, ale pouze funkci (či roli), v jaké vystupují v kognitivním systému, kterého jsou součástí, viz např. Levin (2017). Funkcionalizmus úzce souvisí s ideou mnoha možných realizací [2], viz například Bickle (2016), podle které lze stejné funkce zajistit různý-

[1] Přesněji univerzální Turingovy stroje (Turing, 1936) vykonávající vhodný program.

[2] V originále „multiple realizability".

mi způsoby či prostředky. Spolu s komputacionizmem, tedy tezí, že myšlení je výpočet na vhodných reprezentačních strukturách, viz například Rescorla (2017), tvoří funkcionalizmus a idea mnoha možných realizací základní východiska tradičního přístupu k umělé inteligenci, viz také Havel (2001).

Turingův článek vyvolal řadu reakcí zejména na rozhraní umělé inteligence a filosofie, zmiňme například asi nejznámější z nich, myšlenkový experiment s *Čínským pokojem* (Searle, 1980). Následující sekce stručně shrnou některé z navazujících prací. Pro důkladnější diskuzi *Turingova testu* viz například Tvrdý (2014).

2.2.3 Úplný Turingův test

Původní *Turingův test* však má řadu problémů, a to jak z hlediska snahy být testem inteligence, tak i z hlediska snahy být testem porozumění jazyku. Limitované schopnosti testu vyhodnocovat rozumění jazyku vyplývají z toho, že komunikace mezi strojem a člověkem probíhá přes terminál, tedy v čistě jazykové doméně. Tím je upozaděn vztah jazyka ke světu, který má dvě roviny:

1. Některé jazykové výpovědi reagují na dění ve světě.
2. Jiné jazykové výpovědi způsobují dění ve světě.

Dennett (1980) na základě toho vyvozuje, že *Turingův test* vyhodnocuje jen některé aspekty jazykových schopností. Tohoto si ostatně všímá i Searle (1980), když ve svém myšlenkovém experimentu *Čínský pokoj* upozorňuje na problematiku rozumění jazykovým výpovědím a také na to, že mysli jsou svojí povahou sémantické. Důkladnou analýzu schopnosti *Turingova testu* vyhodnocovat porozumění jazyku podává Schweizer (2012).

Kromě toho, jak upozorňuje Harnad (1991), *Turingův test* ponechává stranou (na jazykovém zprostředkování) i další projevy lidské inteligence v široké paletě lidského chování, které běžně označujeme za inteligentní.

Harnad (1991) ztotožňuje snahu zjistit, zda jsou stroje inteligentní, s otázkou „Mají stroje mysl?“ Tato otázka vede ke starým filosofickým problémům, které se týkají mysli a těla. Jde jednak o metafyzický problém v podobě: „Zda a jaký je vztah mezi mentálními a tělesnými ději?“ viz například Havel (2001), a také o epistemologický a konceptuální problém v podobě otázky: „Mají i jiní lidé (či entity) mysl?“ viz například Hyslop (2014). Harnad (1991) pak považuje řešení problému jiných myslí za relevantní pro test umělé inteligence, přičemž vychází z řešení označovaného jako inference z analogie. Lidé tak podle Harnada přisoudí mysl jiným na základě jejich chování v reálném světě, které je nerozlišitelné od jejich vlastního chování v obdobných situacích.

Proto Harnad (1991) navrhuje tzv. *úplný Turingův test* v následující podobě: Testovaným subjektem je robot přímo interagující s reálným světem a přímo komunikující s lidmi v tomto reálném světě. Předmětem testu jsou jak jazykové schopnosti ve vztahu ke světu, tak i ostatní lidské inteligentní chování. Lze tedy důkladně prověřit, zda robot jazyku rozumí, např. tím, že zvládne vykonat úkol podle jazykových instrukcí, ale i popsat řečí dění ve světě. Descartem uváděnou všestrannost myšlení lze zkoumat i přímo tak, jak se ukazuje v inteligentním jednání.

2.2.4 Skutečně úplný Turingův test

I test rozšířený do podoby *úplného Turingova testu* má však své limity. Ty vyniknou na pozadí filosofické teorie externalizmu, formulované v dílech filosofů Kripkeho (1972), Putna-

ma (1975) a Burgeho (1979). Podle zastánců externalizmu není interní reprezentace dostatečná pro jazykovou referenci, protože ta spoléhá navíc na několik externích aspektů:

1. intersubjektivně přístupný význam daný pravidelnostmi v mikrostoruktuře prostředí,
2. individuální význam získávaný skrze přímé kauzální vztahy s prostředím během osvojování jazyka,
3. dělbu lingvistické práce využívající experty pro přesné vymezení významu
4. a zažitou praxi v sociolingvistickém klanu.

Nezanedbatelnou úlohu při porozumění jazyku tak hraje fakt, že jazyk je společensky, historicky a evolučně vzniklý fenomén.

Na základě sémantického a mentálního externalizmu, tak upozorňuje Schweizer (2012), že jazykové schopnosti ale i inteligentní chování mají druhový rozměr. Pro testy umělé inteligence z toho pak vyplývá, že by se měly zaměřit na schopnosti druhu. Namísto zkoumání jednotlivce v rámci jiného druhu je tak důležité zkoumání jednotlivce v rámci svého druhu.

Schweizer (2012) proto navrhuje tzv. *skutečně úplný Turingův test*, jehož se účastní nikoliv jeden konkrétní subjekt, ale celý druh či kognitivní typ, tedy v případě UI roboti jakožto umělý druh. Za podmínky testu se stále považuje přímá interakce se světem, která má umožnit evoluci komunity. Cílem testu je zjistit, zda si tento kognitivní typ dokáže vyvinout své vlastní inteligentní chování a jazyk. K zodpovězení otázky, zda stroje (či jiné kognitivní typy) mohou myslet, tak nestačí, pokud si stroj osvojí inteligentní chování a jazyk již existující, a tedy pouze parazituje na chování a jazyku jiného druhu (lidí).

Na právě představené linii *Turingova testu* a jeho rozšíření je patrné, že lze nalézt projevy myšlení – inteligence v jazykové komunikaci a jednání ve světě. Důkladné ověření přítomnosti těchto projevů si však žádá provést značně náročný *skutečně úplný Turingův test*, původní Turingova verze je v tomto ohledu nedostatečná. Kromě toho byly naznačeny souvislosti myšlení – inteligence s dalšími schopnostmi, např. učením, či rozuměním. Takto získaný seznam schopností však zřejmě není vyčerpávající a nalezené vztahy nejsou popsány důsledně. Další limity přístupů vycházejících z *Turingova testu* diskutuje také Hernández-Orallo (2000), který zároveň podrobně udává důvody, proč se od těchto přístupů odpoutat.

2.3 Inteligence a další kognitivní schopnosti

Jak bylo naznačeno v předchozí sekci, inteligence není osamocenou vlastností mysli, ale souvisí s dalšími kognitivními schopnostmi. Přesnější vymezení těchto schopností a vztahů mezi nimi provede sekce 2.3.1. V tomto ohledu je také inspirativní oblast tvorby *kognitivních architektur*, kterou se bude zabývat sekce 2.3.2. Přenesením poznatků kognitivních věd do vývoje umělých inteligencí je *psychometrická umělá inteligence*, kterou představí sekce 2.3.3.

2.3.1 Kognitivní paradigma

Ústřední myšlenkou kognitivizmu, jak uvádí de Mey (1992), je to, že zpracování informací nějakým systémem vyžaduje, aby tento systém měl nějakou formu interního modelu či reprezentace prostředí, v němž se vyskytuje. de Mey hovoří o tomto modelu jako o *náhledu na svět* [3] a ukazuje, že nejde o model jediný, ale o relativně volné, proměnné seskupení mnoha různě precizně vypracovaných a svou strukturou odlišných *náhledů na svět*. Tyto modely lze

[3] V originále „world view".

různě kombinovat, přepínat, který z nich bude hrát roli hlavního náhledu, či z nich odvozovat další modely, čímž se vytváří *dynamická kognitivní schémata* – struktury spojující a organizující znalosti o konceptech.

Tyto modely pak mají prostřednictvím formulovaných očekávání dopad na vnímání, které je jimi vždy zprostředkované, a je tedy výsledkem interakce subjektu a objektu. Konkrétní podobu těchto interakcí de Mey (1992) popisuje pomocí *stratifikovaného modelu vnímání*. V tomto modelu přistupuje subjekt k objektu po částečně nezávislých vrstvách a jednotlivé tvary objektu na různých stupních rozlišení mapuje na vztahy v pojmových strukturách svého náhledu na svět, čímž vzniká výsledný amalgám vjemu. Ke vnímání tedy přispívají jak objekt, tak i subjekt, neboť vnímané tvary jsou jak zaregistrované, tak i představované na základě sítě pojmů nesoucí očekávání.

Vnímání spolu s jednáním následně umožňuje získávání znalostí. Těm de Mey (1992) přisuzuje zejména podobu možných interakcí subjektu s objektem uložených ve vhodných strukturách, za které de Mey považuje Minského *rámce*. *Rámce* (Minsky, 1974) jsou komplexní struktury představující poměrně pevnou kostru, do které lze zapasovat konkrétní znalosti na pozice celkem volně spjaté s výchozími hodnotami (tedy s očekáváními). Proces, ve kterém se znalosti formují, lze označit jako *zexplicitňování implicitního* a de Meye zde inspiruje Piagetova *vývojová psychologie* (Piaget, 1936). Jak subjekt interaguje s objektem, učí se nejprve implicitně struktury možných interakcí. Pokud subjekt v interakcích pokračuje, tyto implicitní struktury mohou nabýt explicitní podoby skrze uvědomování si probíhajících interakcí a abstrahování od konkrétního. Od té chvíle pak mohou explicitní znalosti řídit další interakce.

Inteligenci uchopenou jako zpracování informací tak de Mey (1992) zasadil do komplexu kognitivních schopností, zejména reprezentace znalostí a jejich formování skrze interakci se světem pomocí vnímání a jednání, která prochází specifickým vývojem.

2.3.2 Kognitivní architektury

Podobná je i motivace tvorby *kognitivních architektur*. *Kognitivní architektury* jsou doménově nespecifické, tedy obecné, výpočetní modely zachycující esenciální struktury a procesy v mysli. Svou doménovou nespecifičností stojí v protikladu expertních systémů. Tím, že jsou to modely existujících kognitivních systémů, však nejde o umělé inteligence v pravém slova smyslu, mohou však tvorbu umělých inteligencí inspirovat. Díky tomu, že umožňují široce pojatou analýzu chování na mnoha různých úrovních a v mnoha různých doménách, jde však častěji o nástroje k poznání mysli. Viz (Sun, 2007).

Sun (2007) pro jejich tvorbu navrhuje použít *integrovaný hierarchický přístup*. V tomto pojetí různé vědy zkoumají kognici na různých úrovních abstrakce: společenské, psychologické, komponentové a fyziologické. Svými poznatky pak kladou na návrh architektur různá omezení jak zdola, tak i shora. Skrze integraci jednotlivých poznatků různých věd tak lze zaručit vysokou kognitivní realističnost vytvářené architektury.

Sun (2007) uvádí schopnosti, které by *kognitivní architektury* měly implementovat, přičemž se obvykle jejich tvůrci musí rozhodnout, které z nich budou integrální a které sekundární:

- vnímání a kategorizace,
- reprezentace, paměť a odvozování,
- rozhodování a řešení problémů,
- plánování a jednání,

- komunikace,
- metakognice, motivace a cíle,
- učení.

Jako příklady *kognitivních architektur* lze uvést architekturu *CLARION* založenou na duální reprezentaci znalostí oddělující implicitní a explicitní znalosti (Sun, 2007), architekturu *ACT-R* založenou na rozlišení procedurálních a deklarativních znalostí (Anderson et al., 2004), či kognitivní model *Uroboros* založený na rekurzivním algoritmickém procesu (Thomsen, 2013).

2.3.3 Psychometrická umělá inteligence

Jiný přístup čerpá z psychologie, konkrétně z jejího odvětví psychometrie. Psychometrie se zabývá systematickým měřením psychologických vlastností (zejména inteligence) u lidí, ale i u zvířat pomocí různých testů. Bringsjord & Schimanski (2003) tvrdí, že psychometrie odpovídá na otázku, co je to inteligence, čehož by obor umělé inteligence měl využít.

Bringsjord & Schimanski (2003) tedy navrhují, aby umělá inteligence byla pojímána jako *psychometrická*. Cílem umělé inteligence by tak mělo být vytvářet systémy, které budou dosahovat solidních výsledků ve všech zavedených a validovaných testech inteligence a dalších mentálních schopností. Zahrnutím testů dalších mentálních schopností (například kreativity) se Bringsjord & Schimanski (2003) vyhýbají příliš úzkému zaměření na IQ testy. Na základě *psychometrické UI* lze snadno založit inženýrsky výhodný iterativní přístup k tvorbě umělých inteligencí, při kterém se do konstruované entity postupně integrují schopnosti potřebné k vyřešení stále dalších testů inteligence.

Jako přístup k testování umělých inteligencí je však *psychometrická UI* poněkud nepraktická, protože uvažuje otevřenou množinu všech zavedených a validovaných testů. *Psychometrická UI* také explicitně neřeší otázku definování inteligence a přenechává ji psychologii. Bringsjord & Schimanski (2003) toto považují za výhodu, ale jak argumentuje například Besold et al. (2015), je však otázkou, nakolik je možné bez úprav převzít testy postavené na psychologických definicích lidské inteligence. Jejich antropomorfnost může být na škodu. Pro uskutečnění *psychometrické UI* tak zřejmě bude potřeba tyto testy vylepšit a zejména zobecnit, jak navrhuje Besold et al. (2015).

Uchopení tématu inteligence v kognitivních vědách na jedné straně přináší deskriptivní poznatky týkající vymezení a usouvztažnění inteligence vzhledem k dalším kognitivním schopnostem. Popis těchto vztahů může být i značně podrobný a nabývat podoby výpočetních modelů konkrétních kognitivních systémů. Vedle toho se kognitivní vědy zabývají i testováním inteligence u existujících kognitivních systémů. Na příkladu právě představené *psychometrické UI* lze ukázat možnosti a limity přímého přenosu takto získaných poznatků do oblasti vývoje umělých inteligencí.

3 Formální definice inteligence

Dosud představené přístupy často neodpovídají na otázku: „Co je inteligence?“ příliš konkrétně a již vůbec ne formálně. Pokud už přinášejí konkrétnější odpovědi, jsou úzce spjaté s existujícími biologickými systémy, což limituje možnosti jejich přímého uplatnění při konstrukci umělých inteligencí. V nedávné době se však na poli *obecné umělé inteligence* objevily nové snahy směřující k explicitnímu a formálnímu zodpovězení této otázky, které zároveň abstrahují od existujících biologických systémů.

Výsledkem je zejména *definice univerzální inteligence*, kterou se bude zabývat sekce 3.1, ale třeba i pokus o její rozšíření do podoby *definice pragmatické obecné inteligence* popsaný v sekci 3.2.

3.1 Definice univerzální inteligence

Nejprve je třeba vysvětlit důvody, proč je potřeba formální definice inteligence, což provede sekce 3.1.1. Problematiku uchopení přirozené inteligence, z čehož by měla jakákoliv snaha o formální definici vycházet, shrne sekce 3.1.2. Na základě toho je možné přistoupit k vlastní formalizaci univerzální inteligence, čímž se bude zabývat sekce 3.1.3. Aplikaci odvozené definice univerzální inteligence na umělé agenty ukáže sekce 3.1.4. Vlastnosti této definice představí sekce 3.1.5.

3.1.1 Potřeba formální definice inteligence

Inteligence je obtížně uchopitelný abstraktní pojem. Souvisí s dalšími obtížně uchopitelnými pojmy jako učení, znalosti, komunikace, jazyk, myšlení či kreativita. Legg & Hutter (2007b) na základě své rešerše tvrdí, že na obecné úrovni panuje v psychologii v zásadě shoda, byť dílčí otázky zůstávají otevřené.

Tento stav Legg & Hutter (2007b) vidí jako příčinu problémů na poli umělé inteligence. Umělá inteligence totiž kvůli absenci obecně přijímané, přesně formulované definice znamená pro každého něco jiného. Odlišnosti mezi stroji a lidmi pak představují další komplikace. Ve výsledku se nedaří dosáhnout obecné umělé inteligence a výzkum často sklouzává k řešení specifických problémů.

Legg & Hutter (2007b) proto vidí potřebu formální definice inteligence, která by pomáhala směřovat výzkum umělé inteligence a zabraňovala dosavadnímu tápání. Formální definice inteligence by měla zobecnit esenciální vlastnosti lidské inteligence, připouštět i jiné než biologické realizace, vycházet z fundamentálních principů a konečně být objektivní a převeditelná na praktický test.

3.1.2 Přirozená inteligence

Ve snaze o vytvoření formální a precizně formulované definice inteligence Legg & Hutter (2007a) prostudovali řadu definic, teorií a testů zabývajících se inteligencí u lidí a zvířat.

Správně aplikované IQ testy u lidí měří něco důležitého, statisticky stálého a dobře zakládajícího predikce. Nepanuje však shoda, zda jde o inteligenci jako takovou, nějaký její typ, či jen nějaký aspekt. Zůstává také otázka, jak moc jsou IQ testy robustní vzhledem k různým předpojatostem, či kulturním závislostem. Viz (Legg & Hutter, 2007a,b).

Kromě testů u lidí probíhají i snahy o testy inteligence u zvířat. Legg & Hutter (2007a,b) považují tuto oblast za velmi důležitou inspiraci pro testování umělých systémů, neboť při testování zvířat je nutné abstrahovat od lidského pojetí konceptu inteligence jakož i od antropomorfních metod testování. Nelze například testovanému subjektu pomocí jazyka předat instrukce k testu, a proto jsou testy pro zvířata založené obvykle na nějaké formě odměňování. Při testování zvířat také dochází k porovnávání různých druhů, a tak mohou do popředí vystoupit vlastnosti, které při porovnání v rámci jednoho druhu nevyniknou.

Psychologie se kromě vlastních testů inteligence snaží i o tvorbu teorií. Podle rešerše, kterou Legg & Hutter (2007a) provedli, jsou zde patrné dva přístupy. Přístup reprezentovaný zejména Sternbergem a Gardnerem považuje inteligenci za řadu více či méně oddělených a

samostatných schopností, viz (Sternberg, 1984) a (Gardner, 1983). Oproti tomu přístup reprezentovaný například Spearmanem či Cattellem považuje inteligenci za obecnou schopnost, která se projevuje v jiných schopnostech, viz (Spearman, 1927) a (Cattell, 1987).

Při vytváření testů inteligence i psychologických teorií často dochází k formulaci více či méně exaktní definice inteligence. Legg & Hutter (2007a,b) si všímají, co mají tyto definice společného: Inteligence je chápána jako nějaká vlastnost jedince interagujícího s prostředím, problémy či situacemi. Tato vlastnost souvisí se schopností uspět při plnění cílů a dosahování záměrů. Často je kladen důraz na učení a přizpůsobení se nějaké novosti či dosud neznámým aspektům v prostředí.

Legg & Hutter (2007a,b) vyvodili ze společných rysů řady psychologických definic, testů a teorií inteligence následující neformální definici: *„Inteligence měří schopnost agenta dosahovat cílů v mnoha různých prostředích.“*

3.1.3 Formalizace univerzální inteligence

Legg & Hutter (2007b) následně přistoupili k formalizaci výše uvedené pracovní definice. Jak je patrné i z neformálního vyjádření, definice se skládá z několika dílčích bloků:

- **Interakce agenta s prostředím** probíhá po krocích dvěma směry:
  - vjemy zasílané od prostředí k agentu, které se skládají z odměny r_i a pozorování o_i,
  - akce a_i zasílané od agenta k prostředí,

  čímž se utváří interakční sekvence $o_1 r_1 a_1 \cdots o_i r_i a_i$ označovaná také jako historie interakcí.
- **Agent** π je funkce přiřazující historii interakcí nějakou akci. Protože agent nemusí být deterministický, je vhodné popisovat ho jako pravděpodobnostní míru nad prostorem akcí podmíněnou historií interakcí.
- **Prostředí** μ je funkce přiřazující historii interakcí nějakou odměnu a pozorování. Opět je vhodné popisovat prostředí jako pravděpodobnostní míru nad prostorem odměn a pozorování podmíněnou historií interakcí. Protože však je vhodné, aby definice umožňovala odvození testu, který bude probíhat s pomocí počítače, omezují Legg a Hutter množinu prostředí E pouze na takové možné pravděpodobnostní míry, které jsou *Turingovsky vyčíslitelné*. Zde je na místě upozornit, že vyčíslitelnost (a z ní vyplývající omezení) se týká pouze pravděpodobnostní míry, nikoliv prostředí jako takového. Na prostředí lze nahlížet tak, jak pojem prostředí naznačuje, tedy jako na svět, vesmír, či nějakou jeho část, ale i jako na konkrétní úlohu, kterou má agent vyřešit.
- **Míra úspěchu v prostředí** je vyjádřena jako maximalizace očekávaných odměn. Aby definice nepředjímala určité časové preference, a tedy nějakou konkrétní diskontní míru, je suma odměn daných prostředím shora omezená 1 a časová preference řešení (tedy distribuce odměn) zabudovaná v prostředí. Tedy prostředí samo (či úloha sama) určuje, zda je lepší pomalé, ale přesnější, nebo naopak rychlé, ale méně přesné řešení. Míra úspěchu v jednom prostředí je pak vyjádřena následující hodnotovou funkcí:

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

- **Celková míra úspěchu** zohledňuje předpoklad, že úspěch v různých prostředích by měl přispívat k naší představě o inteligenci agenta různou měrou. Legg a Hutter proto rozlišují vliv jednotlivých prostředí pomocí jejich *algoritmické pravděpodobnosti* založené na míře *Kolmogorovovy složitosti* K (Kolmogorov, 1963). Pokud lze pro popis pravděpodobnostní míry prostředí použít krátký program, pak má prostředí nízkou *Kolmogorovovu složitost* (ta je založena na délce nejkratšího programu popisujícího sekvenci bitů) a skrze *algoritmickou pravděpodobnost* velkou váhu. Komplexní prostředí mají tedy menší dopad na celkový výkon agenta než prostředí méně složitá.[4] Celkovou míru úspěchu v prostředích vyjadřuje následující vzorec:

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

přičemž *Kolmogorovova složitost* K řetězce bitů x je dána délkou l nejkratšího programu p, který při spuštění na *Turingově stroji* $\mathcal{U}$ vypíše tuto sekvenci x.

Legg & Hutter (2007b) tedy svou pracovní definici inteligence: *„Inteligence měří schopnost agenta vést si dobře v mnoha různých prostředích,“* formalizovali do podoby uvedené v rovnici 1. Definici popsanou touto rovnicí nazývají *univerzální inteligence*.

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

přičemž *Univerzální inteligence* Υ agenta π je dána jeho schopností dosahovat cílů popsanou hodnotovou funkcí V_μ^π jako očekávanou sumu všech budoucích odměn nad množinou E prostředí μ vážených pomocí *Kolmogorovovy složitosti* K.

3.1.4 Univerzální inteligence různých agentů

Definice univerzální inteligence není praktickým testem, může ale posloužit jako vodítko při vyhodnocování inteligence agentů. K tomu je potřeba odhadnout, vůči jakým typům prostředí si agent povede dobře a také alespoň rámcově, jak dobře si povede. Legg & Hutter (2007b) ukazují uplatnění univerzální inteligence na příkladu několika typů agentů:

- *Náhodný agent* – protože bude své akce vybírat náhodně, neodhalí žádné pravidelnosti v prostředích, a tedy dosáhne jen nízké V_μ^π ve většině prostředí, takže bude mít i nízkou Υ.
- *Specializovaný agent* – agent velmi dobrý v určité úloze (např. šachový program) bude mít v prostředích odpovídajících této úloze vysokou V_μ^π, ale mimo specializaci (ve většině prostředí) nízkou V_μ^π. Celková Υ bude nízká.
- *Obecný jednoduchý agent* – tedy agent schopný vést si statistiku úspěšnosti akcí a vykonávat náhodný průzkum, odhalí základní pravidelnosti prostředí. Měl by tak dosáhnout vyšší V_μ^π alespoň v jednodušších prostředích, a tedy i vyšší Υ než předchozí agenti.

[4] Důvod volby této konstrukce v definici *univerzální inteligence* vyplyne při pohledu z perspektiva agenta, který formuluje hypotézy o prostředí. V souladu s principem Occamovy břitvy tedy jde o preferenci jednodušších hypotéz. Zároveň však agent zvažuje i komplexní hypotézy, pokud jsou konzistentní s dosavadní historií interakcí, což zase respektuje Epikurův princip.

- *Obecný agent s historií* – tedy agent schopný navíc korelace aktuální akce s historií interakcí, odhalí více pravidelností v prostředích (zejména i takové, které bez korelace s historií objevit nelze). Měl by tak dosáhnout vyšší V_μ^π ve vícero prostředích, a tak i vyšší Υ než předchozí agenti.
- *Obecný agent s historií a plánováním* – plánování dále než za aktuální odměnu umožní využít pravidelnosti v prostředích, které bez plánování využít nejde, agent tak dosáhne ještě vyšší V_μ^π ve vícero prostředích, a tak i vyšší Υ než předchozí agenti.
- *Velmi inteligentní agent* – musí dosahovat dobrých výsledků ve většině jednoduchých prostředí, ale i slušných výsledků v mnoha komplexních prostředích. Jeho V_μ^π tedy musí být většinou vysoká. Aby byl agent považován za vysoce inteligentního, musí být nejen velmi úspěšný, ale také velmi obecný.
- *Superinteligence* – tedy agent s maximální Υ, vždy vybere akci s nejvyšší očekávanou budoucí odměnou. Musí být schopna dokonalé predikce, v jakém prostředí se nachází, díky čemuž dosáhne maximální V_μ^π.
- *Člověk* – lze očekávat, že dokáže poznat strukturu jednoduchých prostředí a využít ji pro maximalizaci odměn, u složitějších prostředí je ale nejasné, nakolik je člověk omezen evoluční adaptací (specializací) na určitá prostředí.

Optimální agent vzhledem k definici univerzální inteligence je označován jako AIXI. Podrobněji je AIXI představuje Hutter (2007). Způsob jakým AIXI vybírá své akce je popsán rovnicí 2.

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

kde a_i jsou akce agenta AIXI (a_k je aktuální akce), r_i jsou odměny a o_i pozorování zaslané prostředím a p je program délky l (hypotéza o prostředí) spuštěný na *univerzálním Turingově stroji* $\mathcal{U}$, přičemž vstupem programu je sekvence akcí agenta a výstupem sekvence pozorování a odměn.

Hutter (2012) komentuje rovnici AIXI takto:

> *„Sekvenční teorie rozhodování (Bellmanova rovnice) formálně řeší problém racionálních agentů v nejistých světech za předpokladu, že je skutečná pravděpodobnostní distribuce prostředí známá. Pokud je prostředí neznámé, Bayesiáni nahrazují skutečnou pravděpodobností distribuci vyváženou směsicí distribucí z nějaké třídy hypotéz. Využití velké třídy všech (částečných) měr, které jsou (částečně) vyčíslitelné na Turingově stroji, respektuje Epikurův princip, který učí nezahazovat žádnou (konzistentní) hypotézu. Aby nebyla ignorována Occamova břitva, která vybírá nejjednodušší hypotézu, Solomonoff definoval univerzální distribuci, která přiřazuje vysoké váhy jednoduchým prostředím a nízké váhy složitým prostředím, přičemž složitost se kvantifikuje dle Kolmogorova.“* [5] (Hutter, 2012)

[5] V originále: *„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.“* (Hutter, 2012)

3.1.5 Vlastnosti univerzální inteligence

Univerzální inteligence, kterou předložili Legg & Hutter (2007b), lze nahlížet jako zobecnění dřívější práce týkající se *C-testů* (Hernández-Orallo, 2000) ze statických do dynamických prostředí. Definice Legga a Huttera je pokusem o co nejpřesnější a nejsilnější možnou definici inteligence, i když to přináší komplikace při testování. Konkrétně, využití *Kolmogorovy složitosti*, která není vyčíslitelná, znamená, že ani hodnota Υ není vyčíslitelná. Stejně tak to, že definice uvažuje všechna *Turingovsky vyčíslitelná prostředí*, kterých je nekonečně mnoho, a interakci mezi agentem a prostředím, která jde do nekonečna kroků, představuje problém při převodu do podoby prakticky použitelného testu. Takový test bude muset být založen na aproximacích, které se nějakým způsobem vyrovnají s těmito třemi aspekty nevyčíslitelnosti Univerzální inteligence. [6]

Krom toho, že definice univerzální inteligence není snadno převeditelná na praktický test, má však řadu žádoucích vlastností (Legg & Hutter, 2007b):

- **Validnost** – definuje inteligenci, ne nějakou související kvalitu, nebo jen aspekt inteligence. Legg a Hutter na základě rešerše psychologických poznatků abstrahovali široce přijímanou neformální definici inteligence a formalizovali ji.
- **Smysluplnost** a **informativnost** – dává absolutní porovnatelné měřítko (založené na skalární hodnotě), pomocí nějž lze přirozeně seřadit schopnosti jednoduchých agentů. Zároveň dosažení vysoké Υ předpokládá, že je agent schopen vést si velmi dobře v mnoha jednoduchých i složitějších prostředích, a tedy splňuje očekávání kladená na velmi inteligentní entitu. Definice je snadno uplatnitelná i na nové agenty.
- **Široký rozsah** – umožňující porovnat velmi nízkou inteligenci jednoduchých agentů i velmi vysokou inteligenci superinteligence AIXI.
- **Obecnost** – srovnává výsledky agentů ve všech (vyčíslitelných) prostředích, která navíc mohou být značně odlišná, nikoliv v nějaké specifické úloze.
- **Nepředpojatost** – testové úlohy nejsou výsledkem kulturní nebo jiné předpojatosti, ale závisí pouze na *referenčním Turingově stroji* $\mathcal{U}$.
- **Fundamentálnost** – definice je založena na „pevných základech“ (koncepty výpočtu, informace a složitosti, které stojí v samém středu počítačové vědy).
- **Formálnost** – jasně definovaná matematická rovnice s minimem prostoru pro nejednoznačnost.
- **Objektivnost** – nezávisí na subjektivních kritériích.
- **Univerzálnost** – není antropocentrická.

I když je definice univerzální inteligence nepředpojatá vůči kulturním či jazykovým vlivům, závisí skrze *Kolmogorovovu složitost* na volbě *referenčního Turingova stroje* $\mathcal{U}$, vůči kterému se vyhodnocuje délka nejkratších programů popisujících prostředí. Tato závislost není tak zásadní pro velmi inteligentní agenty jako AIXI, zejména když budou mít dost času na trénink se zvoleným referenčním strojem (Legg & Hutter, 2007b), ale může být důležitá pro

[6] To však Legg & Hutter (2007b) nepovažují za zásadní námitku proti své definici. Pojem založený na konkrétní množině testů je problematický tím, že tyto testy lze často obejít původně nezamýšlenými způsoby. Pak by bylo nutné buď přiznat definovanou vlastnost i tam, kde ve skutečnosti není, nebo pojem redefinovat. Pokud si však budeme vědomi slabiny testů od počátku díky tomu, že nepřesně vystihují ideální definici (naše schopnost testovat danou vlastnost je tedy limitovaná), lze se tomuto problému snáze vyhnout, a to zejména bez potřeby opakovaně redefinovat danou vlastnost.

jednodušší agenty, ba i způsobit zásadní problémy, jak ukazuje Hibbard (2009). Zvolený referenční stroj $\mathcal{U}$ totiž dle Hibbarda zakládá předpojatost vůči početně malé skupině prostředí s krátkými a tedy jednoduchými programy, které pak dominují pomocí definice měřenou inteligenci agentů. Hibbard (2009) zároveň ukazuje, že tuto předpojatost lze podle uvážení redukovat stanovením minimální délky programů popisujících prostředí. Hibbard také ukázal, že při specifické konspiraci mezi zvoleným referenčním strojem, prostředím μ a agentem π lze naměřit vyšší univerzální inteligenci tohoto agenta π, než jaké dosáhne AIXI, avšak pouze za podmínky, že referenční stroj použitý v definici univerzální inteligence k vyhodnocení *Kolmogorovovy složitosti* je jiný, než jaký používá AIXI k simulaci hypotéz o prostředích. Definice univerzální inteligence tedy dle Hibbarda neměla připouštět, aby AIXI používalo odlišný referenční stroj, než jaký používá definice.

Alternativní řešení předpojatosti definice *univerzální inteligence* vůči referenčnímu stroji navrhl Hernández-Orallo (2015, 2017). Jeho přístup je založen na myšlence, že je to ve skutečnosti komplexita řešení, která určuje obtížnost problému (prostředí). Navrhuje proto změnit způsob celkové agregace skóre tak, aby nově zahrnoval tuto myšlenku *obtížnosti prostředí*, kterou Hernández-Orallo definoval jako hodnotu Levinovy Kt složitostní funkce pro nejjednodušší řešení problému.

Hibbard (2009) však také dokázal, že míry inteligence musí být založeny na nerovnoměrně rozložených vahách prostředí. Tuto podmínku Legg-Hutterova definice univerzální inteligence splňuje právě použitím *algoritmické pravděpodobností* ve výrazu pro celkovou míru úspěchu agenta v prostředích.

3.2 Definice pragmatické obecné inteligence

Goertzel (2010) uvádí, že má Legg-Hutterova definice několik nedostatků z hlediska aplikace na reálné agenty v reálných prostředích. Goertzelova kritika, kterou představí sekce 3.2.1, se soustřeďuje do tří oblastí: zacházení s cíli, zacházení s prostředími a neuvažování výpočetních nároků v definici univerzální inteligence. Na základě této kritiky Goertzel navrhuje definici pragmatické obecné inteligence, již představí sekce 3.2.2, a také její verzi započítávající spotřebu výpočetních zdrojů, jak ukáže sekce 3.2.3. Kromě toho se Goertzel věnuje otázce obecnosti versus specifičnosti inteligence, jak naznačí sekce 3.2.4.

3.2.1 Kritika definice univerzální inteligence

Z hlediska způsobu práce s cíli si Goertzel (2010) všímá problematičnosti explicitních cílů definovaných v prostředí a s tím související odměny přicházející z prostředí. Zjevně existují cíle, které si agenti dávají sami, a jen sami agenti vědí, zda je splnili. Navíc zřejmě ne všechno inteligentní chování sleduje nějaké cíle. Goertzelovým řešením je jednak rozšíření definice inteligence o další pravděpodobnostní strukturu popisující cíle, a dále také pohled na tuto definici jako na nepřímo aplikovatelnou na reálné agenty z pozice hypotetických situací. Pro vyhodnocení jejich inteligence je tedy podle Goertzela třeba, aby jiný agent z pozorování chování testovaného agenta vyvodil jeho inteligenci podle upravených měr.

Co se týče prostředí Goertzel (2010) upozorňuje na to, že inteligentní agenti mohou být na některá prostředí adaptovaní, ať už evolučně – jako jsou lidé, nebo návrhem – jako jsou umělé systémy. Pro vyhodnocení inteligence tak nemusí být vždy zajímavá absolutní obecnost (univerzálnost) agenta, ale spíše obecnost, která je vůči některým prostředím předpojatá. Namísto fixace Solomonoff-Levinovy *univerzální pravděpodobnostní distribuce* prostředí v původní definici navrhuje Goertzel uvažovat různé pravděpodobnostní distribuce prostředí. Pro porovnání s lidmi pak může být zejména zajímavá taková pravděpodobnostní distribuce, která

odpovídá rozložení prostředí, ve kterých se lidé běžně vyskytují a inteligentně v nich jednají.

Původní definice neuvažuje žádné aspekty výpočetní náročnosti inteligence, což jak uvádí Goertzel (2010) nereflektuje to, že se reální inteligentní agenti musí vždy vypořádat s omezenými zdroji. Goertzel proto navrhuje uvažovat jednak omezenou paměť agentů přímo v definici, a dále zakomponovat do definice možnost normalizace výsledné inteligence podle pravděpodobnostní distribuce spotřebovaných výpočetních prostředků.

3.2.2 Pragmatická obecná inteligence

Na základě této kritiky Goertzel (2010) zavádí *definici pragmatické obecné inteligence*, jak popisuje rovnice 3:

$$\Pi(\pi) \equiv \sum_{\mu \in E, g \in G, T} \nu(\mu)\gamma(g,\mu)V^{\pi}_{\mu,g,T}, \tag{3}$$

kde *pragmatická obecná inteligence* Π agenta π je dána jako jeho schopnost dosahovat komplexních cílů v komplexních prostředích, jak popisuje hodnotová funkce $V^{\pi}_{\mu,g,T}$ jako očekávanou sumu budoucích odměn relativně k pravděpodobnostní distribuci ν prostředí μ a pravděpodobnostní distribuci γ cílů g v časovém intervalu T.

Goertzel (2010) uvádí mírně modifikovanou hodnotovou funkci $V^{\pi}_{\mu,g,T} \equiv \mathbb{E}\left(\sum_{i=s}^{t} r_g(I_{g,s,i})\right)$, ve které vztahuje úspěšnost agenta k cíli g zadanému na začátku časového intervalu T, v němž agent bere do úvahy očekávané s cílem spjaté odměny r_g ve všech interakčních sekvencích $I_{g,s,i}$ vybraných podle aktuálního prostředí μ.

3.2.3 Efektivní pragmatická obecná inteligence

Aby zohlednil výpočetní nároky inteligence, zavádí Goertzel (2010) *definici efektivní pragmatické obecné inteligence*, jak popisuje rovnice 4:

$$\Pi_{\text{Eff}}(\pi) \equiv \sum_{\mu \in E, g \in G, Q, T} \frac{\nu(\mu)\gamma(g,\mu)\eta_{\pi,\mu,g,T}(Q)}{Q} V^{\pi}_{\mu,g,T}, \tag{4}$$

kde $\eta_{\pi,\mu,g,T}$ udává pravděpodobnost, že agent π v prostředí μ při plnění cíle g v časovém intervalu T spotřebuje výpočetní zdroje Q. Použití pravděpodobnostní distribuce η umožňuje uvažovat nedeterministické agenty, Q je pak pro zjednodušení amalgámem časových, paměťových, energetických a jiných zdrojů a lze ho očekávat někde mezi kladnými reálnými čísly.

Protože pravděpodobnostní distribuce uvažované v Goertzelem upravených definicích nemusejí být Solomonoff-Levinova *univerzální pravděpodobnostní distribuce*, uvedené sumy nemusí konvergovat a zjišťování podmínek, za kterých k tomu dojde, je obtížné. (Goertzel, 2010)

3.2.4 Intelektuální šíře agenta

Goertzel (2010) také uvádí počáteční pokus o vymezení obecnosti versus specifičnosti inteligence, který označuje jako *intelektuální šíři agenta*. Jeho přístup spočívá v možnosti vytvořit fuzzy množinu kontextů, vůči kterým je nějaký agent inteligentní. Za kontext Goertzel označuje trojici: prostředí, cíl a časový interval. Tuto fuzzy množinu pak lze normalizovat do podoby pravděpodobnostní distribuce a zjistit její entropii. Sám autor však uvádí, že tento přístup je limitovaný tím, že neuvažuje vzájemné závislosti mezi prostředími, resp. cíli, vůči kterým jsou agenti inteligentní.

4 Testy inteligence založené na formálních definicích

Představené formální definice inteligence sice podrobně odpovídají na otázku: „Co je to inteligence?“ a to i v podobě, kdy lze jejich odpovědi vztáhnout na umělé systémy, nejsou však prakticky realizovatelnými testy. Přínos explicitně formulovaných formálních definic inteligence však spočívá i v možnosti praktický test z nich odvodit. Správně odvozený test nejen že bude prakticky proveditelný, ale díky svému ukotvení v precizní formální definici má velkou šanci být i skutečně validní.

Sekce 4.1 shrne návrh a prototypovou implementaci *kdykoliv přerušitelného testu inteligence*. Sekce 4.2 pak představí *test algoritmického IQ*.

4.1 Kdykoliv přerušitelný test inteligence

Hernández-Orallo & Dowe (2010) navrhli *kdykoliv přerušitelný test inteligence*,[7] jako test inteligence určený pro současné i budoucí, umělé i biologické agenty, přičemž by tento test měl zvládnout vyhodnotit jak agenty s libovolně velkou (či malou) inteligencí, tak i agenty interagující se světem v libovolných časových měřítcích (krátkých i dlouhých). Test lze kdykoliv přerušit, přičemž vydá tak přesnou aproximaci výsledku, jak mu to dostupný čas na testování umožní. Test spojuje *definici univerzální inteligence* (Legg & Hutter, 2007b) upravenou tak, aby byla vyčíslitelná, s dřívější prací týkající se *C-testů* (Hernández-Orallo, 2000) a *Turingova testu vylepšeného o indukci* (Dowe & Hájek, 1998).

Sekce 4.1.1 shrne, jak návrh testu řeší nevyčíslitelnost definice *univerzální inteligence*. Specifikace *kdykoliv přerušitelného testu inteligence* přináší dva další zajímavé návrhy týkající se adaptivnosti testu a vztahu inteligence a času, které uvede sekce 4.1.2. Sekce 4.1.3 stručně uvede později představenou prototypovou implementaci navrženého testu.

4.1.1 Odstranění nevyčíslitelnosti univerzální inteligence

Hernández-Orallo & Dowe (2010) se vypořádali se třemi aspekty nevyčíslitelnosti *univerzální iteligence* následujícími způsoby:

- Namísto všech prostředí uvažují pouze jejich vzorek, pomocí kterého lze tuto nekonečnou množinu aproximovat. To však vznáší otázku po *diskriminační síle* jednotlivých prostředí v omezeném vzorku, aby do něj vybíraná prostředí co nejvíce přispěla k vyhodnocení inteligence testovaného agenta. Hernández-Orallo a Dowe proto navrhli, aby vzorek zahrnoval pouze prostředí *citlivá vůči odměně*. Agent tak bude testován jen na prostředích, kde záleží na jeho chování, tedy kde jeho chování může vždy skutečně ovlivnit obdržené odměny. Všechna prostředí, která chování agenta zcela či z části ignorují, jsou pak ze množiny prostředí na rozdíl od původní definice vyloučena.
- Test používá omezený počet interakcí mezi agentem a prostředím namísto toho, aby uvažoval nekonečnou interakci. To však otevírá otázku, jak vhodně kombinovat odměny do celkového skóre. Hernández-Orallo a Dowe navrhli zprůměrování odměn počtem interakcí. Kromě toho také požadují, aby se používala pouze *vyvážená* prostředí, kde se odměny vyskytují v intervalu od -1 do $+1$ a kde by náhodné chování agenta vedlo ke skóre v průměru se pohybujícím kolem nuly. Tato omezení jsou předpokladem pro smysluplnost zprůměrovaného skóre.
- Nevyčíslitelnou *Kolmogorovovu složitost* nahradili Hernández-Orallo a Dowe omeze-

[7] V originále „Anytime Intelligence Test“.

nou a vyčíslitelnou složitostní funkcí, která vychází z Levinovy Kt složitostní funkce. Tuto funkci založenou na horním odhadu nejdelší doby výpočtu jedné interakce a omezenou počtem interakcí s prostředím v testu označují jako $Kt^{\max}$. Funkce je tak kromě využití jako distribuce prostředí zachovávající Occamovu břitvu také použita k vynucení časového limitu výpočtu odměn a pozorování při interakci agenta s prostředím.

Referenční stroj použitý v testu je explicitně uveden jako jeho parametr a může zahrnovat velmi omezené stavové automaty, ale i univerzální *Turingovsky úplné* stroje. (Hernández-Orallo & Dowe, 2010)

4.1.2 Čas a inteligence v adaptivním testu

Vedle toho se Hernández-Orallo & Dowe (2010) zaměřili také na dva další aspekty, které považují za důležité u testu inteligence, a sice na *adaptivnost testovacího procesu* a na *vztah mezi časem a inteligencí*.

Fyzikální čas je začleněn do testu jak na straně prostředí, tak na straně agenta. Skrze funkci $Kt^{\max}$ je zajištěn přijatelný limit výpočtu odměny a pozorování prostředím. Dění v prostředí tak lze považovat za okamžité a průběh testu nebude ze strany prostředí zbytečně zdržován. Na straně agenta pak podle Hernándeze-Oralloa a Dowea (2010) nejde zcela oddělit inteligenci a reakční rychlost agenta, takže by čas měl do testu vstupovat buď skrze časový limit testu, nebo i skrze zahrnutí do výsledného skóre. V takovém případě nedochází k průměrování získaných odměn za pevně stanovený počet interakcí, ale tento počet závisí na časovém limitu testu a rychlosti agenta. Metoda průměrující odměny navíc bere do úvahy i prodlevu od poslední provedené akce, čímž zabraňuje agentu, aby oddalováním dalších akcí po získání výhodných odměn vychýlil výsledné skóre.

Proces testování je adaptivní vůči složitosti prostředí a časovému limitu pro test jednoho prostředí, aby se test přizpůsobil jak úrovni inteligenci agenta, tak i časovému měřítku, na kterém agent funguje. Protože Hernández-Orallo & Dowe (2010) chtějí test použitelný na široké spektrum inteligence i rychlosti agentů, stanovují počáteční hodnoty komplexity prostředí co nejnižší a časový limit co nejkratší. Časový limit se postupně prodlužuje, pokud agent nestíhá reagovat. Obtížnost prostředí se zvyšuje v případě dosažení dostatečně vysokých odměn, ale i snižuje v případě dosažení dostatečně nízkých odměn. Zde je důležitá vyváženost obou mechanizmů, aby agent nemohl podvádět, tj. bez negativního dopadu na výsledné skóre zůstávat u lehčích prostředí.

4.1.3 Prototypová implementace testu

Insa-Cabrera et al. (2011) představili prototypovou implementaci zjednodušené verze *kdykoliv přerušitelného testu inteligence* podle návrhu Hernándeze-Oralloa a Dowea (2010) používající jako referenční stroj také poněkud zjednodušenou verzi *nepředpojaté třídy univerzálních prostředí* definované Hernándezem-Oralloem (2010). S touto prototypovou implementací provedli experimenty porovnávající inteligenci lidí s umělým agentem založeným na algoritmu Q-učení (Watkins, 1989). Toto srovnání bylo možné díky myšlence subjektově specifických rozhraní ke stejnému testu. Tato rozhraní mění reprezentaci odměn, akcí a pozorování podle typu testovaného subjektu.

Prototypová implementace *kdykoliv přerušitelného testu inteligence* využívá jednoduchá prostředí založená na stavových prostorech o proměnné velikosti, kterými se agent může pohybovat svými akcemi (Insa-Cabrera et al., 2011). Odměny a tresty jsou generovány dvěma procesy (v terminologii původních autorů označovanými jako agenty), které se pohybují prostředím podle pevně daného vzorce akcí prováděného ve smyčce s náhodným počátečním

bodem. To zaručuje, že prostředí jsou *vyvážená* a *citlivá vůči odměně*. Délka LZ komprimovaného řetězce akcí generujících odměny a tresty se využívá jako odhad složitosti prostředí. Ten je tak proveditelný, avšak značně vzdálený od funkce $Kt^{\max}$ navrhované v *kdykoliv přerušitelném testu inteligence* nebo od *Kolmogorovovy složitosti* v definici *univerzální inteligence*. Testovaný agent má povolen omezený počet interakcí v závislosti na počtu stavů v prostředí. Prototypová implementace však neuvažuje časové měřítko agenta, čímž postrádá jeden z klíčových aspektů původního návrhu. Použitá prostředí navíc nejsou generována *Turingovsky úplným procesem* a jsou plně pozorovatelná agentem. Prototyp tak umožňuje použít pouze značně omezenou množinu prostředí oproti *definici univerzální inteligence*.

4.2 Test algoritmického IQ

Legg & Veness (2013) zkoumali potenciál *definice univerzální inteligence* (Legg & Hutter, 2007b) pro odvození prakticky proveditelného testu, přičemž reflektovali některé z myšlenek obsažených v *kdykoliv přerušitelném testu inteligence* (Hernández-Orallo & Dowe, 2010). Ve výsledku navrhli *test algoritmického IQ* (*AIQ test*) jako pokud možno věrnou aproximaci míry *univerzální inteligence* agenta.

Sekce 4.2.1 shrne, jak byla definice transformována do podoby AIQ testu. Referenční stroj a v něm spouštěné programy prostředí popíše sekce 4.2.2. Agenty, které lze testovat, představí sekce 4.2.3.

4.2.1 Odvození testu algoritmického IQ

Transformace původní definice si vyžádala mnoho kroků, které měly za cíl odstranit tři aspekty nevyčíslitelnosti přítomné v definici univerzální inteligence tak, aby její hodnotu bylo možné aproximovat (Legg & Veness, 2013):

- **Interakce agenta s prostředím** probíhá stejným způsobem jako v případě definice univerzální inteligence. Test však odstraňuje první aspekt nevyčíslitelnosti definice použitím konečného počtu interakcí autory pojmenovaným jako *délka epizody*, kterou zde označím k.
- **Agent** π zůstává stejný jako v definici.
- **Generování vzorku prostředí** musí zachovat ideu Occamovy břitvy a zároveň se vyhnout nevyčíslitelné *Kolmogorovovy složitosti*. Legg a Veness proto volí Solomonoffovu *univerzální distribuci* (Solomonoff, 1964a,b), která přiřazuje pravděpodobnost bitovým sekvencím začínajícím konečným řetězcem x počítaným programem p na *referenčním Turingově stroji* $\mathcal{U}$ následujícím způsobem:

$$M_{\mathcal{U}}(x) := \sum_{p:\mathcal{U}(p)=x*} 2^{-l(p)}.$$

  Kratší program tak má vyšší pravděpodobnost zahrnutí do vzorku v souladu s Occamovou břitvou popsanou *Kolmogorovovou složitostí*.

- **Program prostředí** p je program popisující prostředí μ. Pro test se podle univerzální distribuce na referenčním stroji $\mathcal{U}$ vytváří konečný vzorek S sestávající z N programů prostředí $p_1 \cdots p_N$ stejnoměrným generováním bitů, dokud se nedojde na konec daného programu. Prostředí μ tak může být popsáno vícero programy p_i. Použití konečného vzorku prostředí tak odstraňuje zbývající aspekt nevyčíslitelnosti definice univerzální inteligence.

- **Míra úspěchu v prostředí** je popsána empirickou hodnotovou funkcí $\hat{V}_{p_i}^{\pi}$ jako celková odměna dosažená agentem π za jeden průběh programu prostředí $\mathcal{U}(p_i)$ zprůměrovaná počtem interakcí k. Odměny přidělené prostředím nejsou shora omezené 1 jako u původní definice a nejsou nijak diskontovány pro vyjádření časové preference na této úrovni.
- **Celková míra úspěchu** již nemusí zohledňovat Occamovu břitvu, protože ta je zabudována ve způsobu, jakým je generován vzorek programů prostředí S. Jde tedy o prosté zprůměrování výsledků agenta ve všech N testovaných programech prostředí.

Legg & Veness (2013) tedy transformovali definici *univerzální inteligence* popsanou rovnicí 1 do podoby zachycené v rovnici 5. Tuto aproximaci nazývají *algoritmickým IQ*.

$$\hat{\Upsilon}(\pi) := \frac{1}{N}\sum_{i=1}^{N}\hat{V}_{p_i}^{\pi}, \text{ kde} \qquad \hat{V}_{p_i}^{\pi} := \frac{1}{k}\sum_{i=1}^{k} r_i, \tag{5}$$

přičemž *AIQ odhad univerzální inteligence* $\hat{\Upsilon}$ agenta π je dán jeho schopností dosahovat cílů popsanou empirickou hodnotovou funkcí $\hat{V}_{p_i}^{\pi}$ jako průměrná odměna dosažená agentem za jeden průběh programu prostředí p_i o k interakcích na konečném vzorku N programů prostředí.

4.2.2 Programy prostředí a referenční stroj BF

Program prostředí v AIQ testu je na rozdíl od prostředí použitých v prototypové implementaci *kdykoliv přerušitelného testu inteligence* (Insa-Cabrera et al., 2011) *Turingovsky úplný* program. Tento program počítá aktuální odměnu a pozorování na základě interakční sekvence (Legg & Veness, 2013).

Stejně jako v případě definice *univerzální inteligence* ovlivňuje výběr *referenčního Turingova stroje* (tedy jazyka programů prostředí), které třídy prostředí mají vyšší pravděpodobnost výběru do vzorku S, a tedy činí test závislým na této volbě (Legg & Veness, 2013). Test ve snaze minimalizovat tuto závislost používá poměrně jednoduchý BF referenční stroj (Müller, 1993), který Legg a Veness rozšířili o instrukci pro konec programu a pro zápis náhodného symbolu. To umožňuje nedeterminizmus programů prostředí.

V souladu s požadavky na *vyváženost prostředí* (Hernández-Orallo & Dowe, 2010) jsou spočtené odměny normalizovány do intervalu $< -100; +100 >$, což také ohraničuje nejmenší a nejvyšší AIQ. Způsob, jakým test probíhá, zajišťuje, že AIQ náhodně se chovajícího agenta je 0 (Legg & Veness, 2011, 2013).

Symboly akcí a pozorovaní stejně jako interní stavy prostředí jsou celočíselné hodnoty typu modulo. Stroj používá jednosměrnou vstupní pásku, která je pouze pro čtení. Na této pásce je umístěna aktuální akce agenta a historie nejvýše 24 dalších akcí. Pracovní páska je obousměrná, pro čtení i pro zápis, a má délku 100 000 políček oběma směry. Výstupní páska je pouze pro zápis a obsahuje jedno políčko pro odměnu a nastavitelný počet políček pro pozorování. Tento návrh činí prostředí, která nejsou plně pozorovatelná, velmi pravděpodobnými. BF jazyk používá 10 instrukcí (Legg & Veness, 2011, 2013; Müller, 1993):

- `+` `-` pro inkrementaci, resp. dekrementaci symbolu na pracovní pásce,
- `,` pro čtení aktuálního symbolu vstupní pásky, jeho zápis na políčko pracovní pásky a přetočení vstupní pásky o políčko,

- `.` pro zápis aktuálního symbolu pracovní pásky na aktuální políčko výstupní pásky a přetočení výstupní pásky o políčko,
- `<>` pro posun pracovní pásky vlevo nebo vpravo,
- `[ ]` pro spuštění cyklu, pokud je aktuální políčko pracovní pásky nenulové, respektive pro vymezení konce cyklu,
- `%` pro zápis náhodného symbolu do aktuálního políčka pracovní pásky,
- `#` pro konec programu.

Jako způsob vyrovnání se s problémem zastavení, který je z praktického hlediska shodný s problémem dlouho běžících programů, je výpočet každé interakce limitován 1 000 kroků. Toto je dále zesíleno ukončením programu prostředí v okamžiku, kdy se snaží zapsat více než nastavený počet symbolů odměn a pozorování. Kromě vyloučení dlouho běžících programů ze vzorku S, je i zastoupení neinteragujících programů prostředí (v terminologii Legga a Venesse pasivních programů) značně sníženo povinným výskytem instrukce pro čtení i pro zápis a také vynecháním programů, které vrací konstantní odměny, (Legg & Veness, 2011, 2013). Díky tomu je požadavek na vyloučení *prostředí bez diskriminační síly* (Hernández-Orallo & Dowe, 2010) částečně splněn.

(Legg & Veness, 2011, 2013) používají řadu technik redukujících rozptyl a zrychlujících proces konstrukce odhadu AIQ. S programy prostředí je nejtěsněji spjata metoda stratifikovaného vzorkování, pro kterou se programy prostředí klasifikují do 20 navzájem výlučných vrstev. Deset těchto vrstev používá jako klíč ke klasifikaci přítomnost jednoduchých vzorů v navracených odměnách. Zbývajících 10 vrstev je rozděleno podle délky programu. Během testu se pak pro každého testovaného agenta vybírají programy tak, aby maximálně minimalizovaly rozptyl odhadu jeho AIQ v dané vrstvě. Ve výsledku se tak vybere více programů z takové vrstvy, kde se výsledky agenta mezi jednotlivými programy prostředí nejvíce liší.

Legg-Venessův *test algoritmického IQ* je dostupný jako Open Source prototypová implementace v Pythonu z (Legg & Veness, 2011). Test lze do určité míry konfigurovat: Zejména je možné nastavit počet programů prostředí N a tím ovlivnit přesnost odhadu *AIQ skóre*. Dále je možné určit počet interakcí k mezi agentem a prostředím (délku epizody) a tím ovlivnit, kolik „času“ má agent pro učení, což při dostatečně vysokém nastavení umožní konvergenci skóre agenta. Kromě toho lze měnit počet symbolů pro pozorování a akce a tím ovlivňovat komplexitu prostoru interakcí. Lze také zvýšit počet pozorování, která mohou být vrácena na konci každé interakce, a tak dále ovlivnit komplexitu interakčního prostoru a také potenciálně prodloužit dobu výpočtu programu prostředí skrze z toho vyplývající navýšení limitu pro počet zápisů. (Legg & Veness, 2011, 2013)

4.2.3 Agenty v testu algoritmického IQ

Implementace testu dostupná z (Legg & Veness, 2011) umožňuje testovat agenty dodané v podobě interní implementace nebo v podobě externího programu, který komunikuje s testem prostřednictvím specifického rozhraní. Test v současné verzi zahrnuje implementace následujících jednoduchých agentů: *random*, *freq*, Q_λ subsumující Q_0 a HLQ_λ. Dále je součástí testu rozhraní pro externí implementaci Monte Carlo aproximace agenta AIXI (*MC-AIXI*). Uvedené agenty se liší svými schopnostmi a konfigurovatelností:

- Agent *random* vybírá své akce náhodně a není nastavitelný.
- Agent *freq* provádí ϵ-hladový výběr akce, tedy s konfigurovatelnou pravděpodobností

ϵ vybere náhodnou akci, jinak vybere akci s dosud nejvyšší průměrnou odměnou.

- Agent Q_λ provádí Q-učení se stopami způsobilosti. To funguje pomocí učení akční hodnotové funkce, která vrací očekávanou utilitu volby dané akce v daném stavu stavového prostoru. Agent má následující parametry (Watkins, 1989):
  - počáteční Q hodnotu,
  - diskontní míru stop způsobilosti λ určující nakolik by měly dříve provedené akce ovlivnit učení, při $\lambda = 0$ provádí agent běžné Q-učení, kde záleží pouze na poslední provedené akci,
  - rychlost učení α, při $\alpha = 0$ se Q hodnoty neaktualizují, se zvyšujícími hodnotami α se Q hodnoty aktualizují rychleji, a tedy se agent učí rychleji,
  - pravděpodobnost výběru náhodné akce ϵ pro ϵ-hladový výběr akce,
  - diskontní míru γ určující důležitost budoucích odměn. Při $\gamma = 0$ záleží pouze na aktuální odměně, pro vyšší γ roste role budoucích.
- Agent HLQ_λ provádí vylepšené Q-učení s automaticky se měnící rychlostí učení. Má následující parametry (Hutter & Legg, 2007):
  - mód výběru akcí umožňující zapnout (0) nebo vypnout ϵ-hladový výběr akce,
  - počáteční Q hodnotu,
  - diskontní míru stop způsobilosti λ,
  - pravděpodobnost výběru náhodné akce ϵ,
  - diskontní míru γ.
- Agent *MC-AIXI* je aproximací z hlediska definice univerzální inteligence optimálního agenta AIXI. Rozhraní, které je součástí testu, zpřístupňuje následující parametry agenta (Veness et al., 2011):
  - počet Monte Carlo simulací ovlivňující predikční sílu modelu,
  - hloubka stromu kontextů ovlivňující velikost agentova modelu,
  - horizont prohledávání ovlivňující počet interakcí, pro které model predikuje svá očekávání,
  - průzkum, tedy vlastně ϵ,
  - útlum průzkumu exponenciálně snižující hodnotu ϵ. [8]

Představené agenty zastupují techniky posilovaného učení [9] (Sutton & Barto, 1998). Tyto techniky mají široké spektrum uplatnění. Již Sutton & Barto (1998) uvádějí jednoduché hry jako backgammon a piškvorky či řízení fyzických robotů, ale také problémy jako dispečink výtahů a dynamickou alokaci kanálů v telekomunikacích. Hutter & Legg (2007) ukazují aplikaci svého algoritmu na úloze prostorové navigace ve větrném prostředí. Veness et al. (2011) demonstrují svého agenta na několika jednoduchých bludištích a hrách jako je Kuhnův poker

[8] Tento parametr není zpřístupněn v rozhraní zveřejněném v (Legg & Veness, 2011), ale byl používán při experimentech reportovaných v (Legg & Veness, 2013), jak odhalil můj dotaz na autory. Proto jej zde uvádím.

[9] V originále „reinforcement learning“.

proti oponentu hrajícímu podle Nashovy strategie a částečně pozorovatelné verzi známé hry Pacman. Využití sofistikovanějšího agenta vycházejícího z posilovaného učení ukázali Mnih et al. (2015) na příkladu řady her pro Atari.

5 Diskuze a závěr

V tomto článku jsem se snažil perspektivou *obecné umělé inteligence* ukázat, jak se vyvíjely přístupy k vyhodnocování inteligence umělých systémů. Tím byl v předchozích sekcích splněn první cíl tohoto článku. Nyní je možné přistoupit k posouzení jednotlivých přístupů, mezi kterými lze identifikovat určité společné rysy, na jejichž základě lze vyčlenit dvě principiálně odlišné skupiny přístupů:

1. První skupina, jejímž asi nejznámějším zástupcem je *Turingův test* (Turing, 1950), staví na přístupu, který bych označil jako *testování komplexních projevů inteligence*. Tyto přístupy identifikují nějakou schopnost nebo množinu schopností, které považujeme za úzce spjaté s inteligencí u lidí (např. schopnost vedení smysluplné konverzace v případě *Turingova testu*). Na základě úspěšného předvedení této schopnosti v nějakém testu pak lze strojům přiznat inteligenci. Debaty se zde zejména vedou o to, jaké schopnosti do takového testu zahrnout. Názornou ilustrací jsou rozšířené verze *Turingova testu*, ale i přístupy vycházející z kognitivní vědy, které akcentují souvislosti inteligence s řadou kognitivních schopností. Nevyřčeným předpokladem těchto přístupů je však to, že *očekávají, že úspěch v této komplexní schopnosti znamená samozřejmě i úspěch v řadě jiných triviálnějších schopností*. Že něco takového nastane vždy a nutně, však považuji za omyl.

2. Druhou skupinu reprezentuje zejména definice *univerzální inteligence* (Legg & Hutter, 2007b). Podle mého názoru velmi důležitou myšlenkou ukrytou v definici *univerzální inteligence* je právě způsob, na základě kterého definice usuzuje na inteligenci umělých systémů. Při posuzování inteligence lidí máme tendenci přisoudit vysokou inteligenci v případě úspěchu v obzvlášť obtížném úkolu, či zvládnutí náročné činnosti. Oprávněnost této tendence spočívá v tom, že jde o lidi, a tedy víme, či předpokládáme, že jsou také inteligentní – ve skutečnosti jen posuzujeme míru inteligence. Situace však není stejná, pokud posuzujeme něco natolik odlišného jako jsou počítačové systémy. Jak ukázaly případy specifických umělých inteligencí, například šachových programů, úspěch v náročném úkolu nemusí nutně znamenat zvládnutí řady jednodušších činností (například hrát piškvorky nejsou šachové programy schopné, a to si navíc tyto úlohy jsou v mnoha ohledech dost podobné). Zde nastupuje definice *univerzální inteligence*, která *přisoudí vysokou míru inteligence entitě úspěšné v náročné činnosti jen v případě, kdy je tato entita zároveň prokazatelně úspěšná i v řadě jednodušších činností*.

S touto perspektivou je také třeba hledět na představenou kritiku definice *univerzální inteligence* a z ní vycházejících přístupů. Pokud například Hibbard (2009) navrhuje zcela vyloučit testování agentů na jednoduchých úlohách, řeší tím sice problém dominance míry inteligence jednoduchými prostředími, ale zároveň v definici opět potlačuje explicitní testování úspěšnosti v jednoduchých prostředích jako způsob ověření skutečné obecnosti agenta. To považuji za významný nedostatek Hibbardovy kritiky. Při omezování dominance míry inteligence jednoduchými prostředími tedy zřejmě bude potřeba sáhnout k nějakému kompromisu. Alternativní řešení, které předložil Hernández-Orallo (2015), je na rozdíl od *složitosti úloh* jakožto obecné vlastnosti pocházející z prostředí založené na konceptu *obtížnosti úloh* jakožto intersubjektivní vlastnosti utvářené z řešení agentů. Hernández-Orallo sice předpokládá test agentů jak na jednoduchých, tak i na obtížných prostředích, provedená změna perspektivy

je však natolik zásadní, že bez další analýzy nelze rozhodnout, nakolik jeho řešení odpovídá výše diskutované ideji za definicí univerzální inteligence.

Úprava definice, kterou navrhuje Goertzel (2010), je však trochu jiného charakteru. Důkladnost testu obecnosti agenta sice omezuje, avšak s argumentem, že někdy je spíše než čirá univerzálnost důležitá nějak předpojatá obecnost, nemohu nesouhlasit. Goertzel se zde dotýká toho, že vyhodnocování inteligence umělých systémů nemůže být odtrženo od reálných úloh, které takové systémy mají řešit. Naráží jednak na problematiku *validace přístupů k vyhodnocování inteligence umělých systémů*, ale i na otázku *relevantnosti metod vyhodnocování obecné inteligence pro konkrétní úlohy* či jejich úzce vymezené skupiny. Přínos Goertzelovy kritiky je pak také v tom, že naznačuje jak vyhodnocovat obecnost inteligence, což je charakteristika, kterou definice univerzální inteligence zahrnuje pouze implicitně.

Přístupy vycházející z definice *univerzální inteligence* se jeví jako vhodné k dalšímu prozkoumání. Návrh *kdykoliv přerušitelného testu inteligence* (Hernández-Orallo & Dowe, 2010) doplňuje původní definici o zajímavé teoretické (vztah inteligence a času), ale i praktické (adaptivnost testovacího procesu) aspekty. *Test algoritmického IQ* (Legg & Veness, 2013) zůstává oproti tomu věrnější původní definici. Kritéria výběru vhodného testu pro další výzkum však jsou i ryze praktického rázu. Ačkoliv tedy souhlasím s tím, že *návrh kdykoliv přerušitelného testu inteligence* je v některých aspektech propracovanější než *test algoritmického IQ*, je však *prototypová implementace kdykoliv přerušitelného testu inteligence* natolik zjednodušená (ty nejdůležitější přínosy návrhu vůbec neimplementuje), že ji *test algoritmického IQ* značně překonává. I proto jsem si jako test pro bližší empirický výzkum vybral právě *test algoritmického IQ*. Tím je splněn druhý cíl tohoto článku. Zde je třeba opětovně zdůraznit, že na základě provedené rešerše lze ke vhodnosti *AIQ testu* jakožto obecné metody pro vyhodnocování inteligence umělých systémů vznést několik závažných připomínek:

- Jak si všímají již Legg & Hutter (2007b) a jak dále analyzuje Hibbard (2009) a také Hernández-Orallo (2015), definice *univerzální inteligence* a tedy i výsledky *AIQ testu* závisejí na výběru *referenčního Turingova stroje*.
- Definice *univerzální inteligence* uvažuje úspěšnost agenta ve všech vyčíslitelných prostředích. Tato množina nutně zahrnuje, jak správně upozorňují Hernández-Orallo & Dowe (2010), mnohá prostředí, která *nemají diskriminační sílu*, a to včetně prostředí tzv. typu „nebe“ či „peklo“. Zatímco tato skutečnost nemusí být pro definici inteligence příliš podstatná, její závažnost v případě praktického testu je kritická, neboť pro provedení testu jsou k dispozici pouze omezené zdroje, které by byly prostředími bez diskriminační síly plýtvány, a test také prověřuje agenta pouze na vzorku prostředí, který tak může být prostředími bez diskriminační síly vychýlen. *Test algoritmického IQ* není proti takovým případům zabezpečen tak dobře jako návrh *kdykoliv přerušitelného testu inteligence*.
- Míra *univerzální inteligence* stejně jako odvozené AIQ skóre zahrnuje některé aspekty inteligence pouze implicitně, zatímco jiné nezahrnuje vůbec. Ačkoliv skóre spojuje míru úspěšnosti agenta v prostředích s mírou jeho obecnosti, kterou Goertzel (2010) označuje jako *intelektuální šíři*, žádný z těchto aspektů není explicitní, což ztěžuje důkladné porovnání agentů. Kromě toho například aspekt efektivnosti systému není zahrnut vůbec, ač jeho zahrnutí obhajuje např. Goertzel (2010), a stejně tak není zahrnut ani aspekt časovosti, který navrhují zahrnout Hernández-Orallo & Dowe (2010).

Limity představeného článku lze spatřovat především v jeho rešeršní povaze. Výběr kandidáta na vhodnou metodu vyhodnocování obecné inteligence umělých systémů, který jsem

provedl, tak není založen na empirickém porovnání jednotlivých metod. Stejně tak výše uváděné připomínky ke vhodnosti *testu algoritmického IQ* vycházejí z provedené rešerše a nebyly empiricky ověřeny. Provedená rešerše je snad vzhledem k cílům článku dostatečně široká, jistě však není zcela vyčerpávající. Ač v článku uplatňuji multidisciplinární pohled a zmiňuji filosofické aspekty daného tématu, jde o sekundární perspektivy, jejichž zpracování by jistě mohlo být prohloubeno.

Uvedené limity článku lze odstranit či umenšit v rámci navazující práce. Ta by se měla soustředit zejména na dvě následující oblasti:

1. Důkladné empirické vyhodnocení *testu algoritmického IQ*. V rámci něj se lze dále zaměřit na:

   (a) potvrzení, či vyvrácení existence (případně určení skutečného rozsahu a závažnosti) v této rešerši identifikovaných limitů testu;

   (b) hledání dalších limitů testu a možností jeho vylepšení.

2. Vylepšení *testu algoritmického IQ* s cílem odstranit či minimalizovat identifikované limity testu a tím jej dále přiblížit do podoby vhodného testu obecné inteligence umělých systémů.

Bude-li k dispozici vhodný test obecné inteligence umělých systémů, je zajímavou otázkou (ač již nad rámec tohoto článku), co s takovým testem půjde dělat? Stručně načrtnu několik možností:

1. Přímočarým použitím vhodného testu je vyhodnocování konkrétních systémů. Jde tedy o porovnávání jednotlivých uměle inteligentních systémů a jejich konfigurací mezi sebou či zjišťování, zda je implementované vylepšení určitého systému skutečně přínosné pro jeho inteligenci. Toto lze považovat za primární cíl a motivaci vývoje evaluačních metod pro umělé systémy.

2. Pokud bude předchozí provedeno na dostatečném množství různých systémů, otevírá se cesta k induktivním generalizacím o skupinách či třídách umělých systémů. Tyto skupiny lze konstruovat například dle toho, ze kterého paradigmatu umělé inteligence systém vychází. To v ideálním případě umožní empiricky odpovídat na řadu obecných otázek na rozhraní teoretické umělé inteligence a filosofie.

3. Vlastní vyhodnocování inteligence umělých systémů se však nemusí provádět pouze externě, může jej provádět i interně sám uměle inteligentní systém. V takovém případě půjde vlastně o určitou formu metaučení, která může být zajímavá jak u sebezdokonalujících systémů, tak i u systémů s velkým množstvím parametrů, které je potřeba konfigurovat. Takový systém by si tedy mohl pomocí vhodného testu vyhodnotit, které z jeho konfigurací jsou z hlediska obecné inteligence vhodnější, a pak se tímto způsobem nakonfigurovat.

Nastíněné možnosti použití reálně znamenají potřebu provádět velké množství testů inteligence. Je tedy nutné, aby použitý test byl co možná nejvhodnější. Vhodnost testu musí být naplněna z mnoha hledisek, počínaje tím, že takový test skutečně měří inteligenci, konče výpočetní efektivitou provádění testu. Míra splnění těchto kritérií nemůže být vyhodnocena jen rešerší, musí již být empiricky ověřena.

Seznam použité literatury

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060. doi: 10.1037/0033-295x.111.4.1036.

Besold, T., Hernández-Orallo, J., & Schmid, U. (2015). Can machine intelligence be measured in the same way as human intelligence? *KI – Künstliche Intelligenz*, *29*(3), 291–297. doi: 10.1007/s13218-015-0361-4.

Bickle, J. (2016). Multiple realizability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Retrieved November, 20, 2017, from https://plato.stanford.edu/archives/spr2016/entries/multiple-realizability/.

Bringsjord, S. & Schimanski, B. (2003). 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)*, (pp. 887–893). Acapulco: IJCAI.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*(1), 73–121. doi: 10.1111/j.1475-4975.1979.tb00374.x.

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. New York: Elsevier.

de Mey, M. (1992). *The cognitive paradigm*. Chicago and London: University of Chicago Press. doi: 10.1007/978-94-009-7956-7.

Dennett, D. C. (1980). The milk of human intentionality. *Behavioral and Brain Sciences*, (3), 428–430. doi: 10.1017/S0140525X0000580X.

Descartes, R. (1637), [1992]. *Rozprava o metodě*. Praha: Svoboda.

Dowe, D. L. & Hájek, A. R. (1998). A non-behavioural, computational extension to the Turing test. In Selvaraj, H. & Verma, B. (Eds.), *Proceedings of International Conference on Computational Intelligence & Multimedia Applications (ICCIMA'98), Gippsland, Australia*, (pp. 101–106). Singapore: World Scientific.

Gardner, H. (1983). *Frames of mind: Theory of multiple intelligences*. New York: Basic Books.

Goertzel, B. (2010). Toward a formal characterization of real-world general intelligence. In Baum, E., Hutter, M., & Kitzelmann, E. (Eds.), *Proceedings of the 3rd International Conference on Artificial General Intelligence (AGI 2010), Lugano, Switzerland*, (pp. 19–24). Amsterdam-Beijing-Paris: Atlantis Press. doi: 10.2991/agi.2010.17.

Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, *5*(1), 1–48. doi: 10.2478/jagi-2014-0001.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, *1*(1), 43–54. doi: 10.1007/BF00360578.

Havel, I. M. (2001). Přirozené a umělé myšlení jako filozofický problém. In V. Mařík, O. Štěpánková, & J. Lažanský (Eds.), *Umělá inteligence 3* (pp. 17–75). Praha: Academia.

Hernández-Orallo, J. (2000). Beyond the Turing test. *Journal of Logic, Language and Information*, *9*(4), 447–466. doi: 10.1023/A:1008367325700.

Hernández-Orallo, J. (2010). A (hopefully) unbiased universal environment class for measuring intelligence of biological and artificial systems. In Baum, E., Hutter, M., & Kitzelmann, E. (Eds.), *Proceedings of the 3rd International Conference on Artificial General Intelligence (AGI 2010), Lugano, Switzerland*, (pp. 182–183). Amsterdam-Beijing-Paris: Atlantis Press. doi: 10.2991/agi.2010.18.

Hernández-Orallo, J. (2015). C-tests revisited: Back and forth with complexity. In Bieger, J., Goertzel, B., & Potapov, A. (Eds.), *Proceedings of the 8th International Conference on Artificial General Intelligence (AGI 2015), Berlin, Germany*, (pp. 272–282). Berlin: Springer. doi: 10.1007/978-3-319-21365-1_28.

Hernández-Orallo, J. (2017). *The measure of all minds*. Cambridge: Cambridge University Press. doi: 10.1017/9781316594179.

Hernández-Orallo, J. & Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, *174*(18), 1508–1539. doi: 10.1016/j.artint.2010.09.006.

Hibbard, B. (2009). Bias and no free lunch in formal measures of intelligence. *Journal of Artificial General Intelligence*, *1*(1), 54–61. doi: 10.2478/v10229-011-0004-6.

Hutter, M. (2007). Universal algorithmic intelligence: A mathematical top→down approach. In B. Goertzel & C. Pennachin (Eds.), *Artificial General Intelligence* (pp. 227–290). Berlin: Springer. doi: 10.1007/978-3-540-68677-4_8.

Hutter, M. (2012). One decade of universal artificial intelligence. In P. Wang & B. Goertzel (Eds.), *Theoretical Foundations of Artificial General Intelligence* (pp. 67–88). Paris: Atlantis Press. doi: 10.2991/978-94-91216-62-6_5.

Hutter, M. & Legg, S. (2007). Temporal difference updating without a learning rate. In Platt, J. C., Koller, D., Singer, Y., & Roweis, S. T. (Eds.), *Proceedings of the 21st Annual Conference on Advances in Neural Information Processing Systems (NIPS 2007), Vancouver, Canada*, (pp. 705–712). New York: Curran Associates, Inc.

Hyslop, A. (2014). Other minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Retrieved November, 20, 2017, from http://plato.stanford.edu/archives/spr2014/entries/other-minds/.

Insa-Cabrera, J., Dowe, D. L., España-Cubillo, S., Hernández-Lloreda, M. V., & Hernández-Orallo, J. (2011). Comparing humans and AI agents. In Schmidhuber, J., Thórisson, K. R., & Looks, M. (Eds.), *Proceedings of the 4th International Conference on Artificial General Intelligence (AGI 2011), Mountain View, USA*, (pp. 122–132). Berlin: Springer. doi: 10.1007/978-3-642-22887-2_13.

Kolmogorov, A. N. (1963). On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, *4*(25), 369–376. doi: 10.1016/S0304-3975(98)00075-9.

Kripke, S. A. (1972). *Naming and necessity*. Cambridge: Harvard University Press.

Legg, S. & Hutter, M. (2007a). A collection of definitions of intelligence. In B. Goertzel & P. Wang (Eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (pp. 17–24). Amsterdam: IOS Press.

Legg, S. & Hutter, M. (2007b). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*(4), 391–444. doi: 10.1007/s11023-007-9079-x.

Legg, S. & Veness, J. (2011). AIQ: Algorithmic intelligence quotient [source codes]. Retrieved June, 26, 2017, from https://github.com/mathemajician/AIQ.

Legg, S. & Veness, J. (2013). An approximation of the universal intelligence measure. In D. L. Dowe (Ed.), *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence* (pp. 236–249). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-44958-1_18.

Levin, J. (2017). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Retrieved November, 20, 2017, from https://plato.stanford.edu/archives/win2017/entries/functionalism/.

Levy, D. & Newborn, M. (1991). *How computers play chess*. New York: Computer Science Press. doi: 10.1007/978-3-642-85538-2_2.

Minsky, M. (1974). A framework for representing knowledge. Technical report. Retrieved November, 20, 2017, from http://web.media.mit.edu/~minsky/papers/Frames/frames.html.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. doi: 10.1038/nature14236.

Müller, U. (1993). dev/lang/brainfuck-2.lha in Aminet. Retrieved June, 26, 2017, from http://aminet.net/package.php?package=dev/lang/brainfuck-2.lha.

Piaget, J. (1936). *Origins of intelligence in the child*. London: Routledge & Kegan Paul.

Putnam, H. (1975). *Mind, language and reality*, chapter The Meaning of 'Meaning', (pp. 215–271). Cambridge: Cambridge University Press.

Rescorla, M. (2017). The computational theory of mind. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford: Metaphysics Research Lab, Stanford University. Retrieved November, 20, 2017, from https://plato.stanford.edu/archives/spr2017/entries/computational-mind/.

Schweizer, P. (2012). The externalist foundations of a truly total Turing test. *Minds and Machines*, *22*(3), 191–212. doi: 10.1007/s11023-012-9272-4.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, (3), 417–457. doi: 10.1017/S0140525X00005756.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm.

Solomonoff, R. J. (1964a). A formal theory of inductive inference, part 1. *Information and Control*, *7*(1), 1–22. doi: 10.1016/S0019-9958(64)90131-7.

Solomonoff, R. J. (1964b). A formal theory of inductive inference, part 2. *Information and Control*, *7*(2), 224–254. doi: 10.1016/S0019-9958(64)90131-7.

Spearman, C. E. (1927). *The abilities of man, their nature and measurement*. New York: Macmillan.

Sternberg, R. J. (1984). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge: Cambridge University Press.

Sun, R. (2007). The importance of cognitive architectures: An analysis based on CLARION. *Journal of Experimental & Theoretical Artificial Intelligence*, *19*(2), 159–193. doi: 10.1080/09528130701191560.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press. doi: 10.1016/S0925-2312(00)00324-6.

Thomsen, K. (2013). The cerebellum in the Ouroboros model, the "interpolator hypothesis". In Shimizu, S. & Bossomaier, T. (Eds.), *Proceedings of the 5th International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE 2013), Valencia, Spain*, (pp. 37–41). Wilmington: IARIA.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2*(42), 230–265.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Tvrdý, F. (2014). *Turingův test: Filozofické aspekty umělé inteligence*. Praha: Togga.

Veness, J., Ng, K. S., Hutter, M., Uther, W., & Silver, D. (2011). A Monte Carlo AIXI approximation. *Journal of Artificial Intelligence Research*, *40*(1), 95–142. doi: 10.1613/jair.3125.

Watkins, C. (1989). *Learning from delayed rewards*. PhD thesis, University of Cambridge, Kings College, Cambridge.

The article has been reviewed. | **Received:** 29 January 2018 | **Accepted:** 29 May 2018
Academic Editor: Stanislava Mildeova