Filosofická fakulta Univerzity Palackého v Olomouci

Katedra filosofie

Kurt Gödel a problematika umělé inteligence

DIPLOMOVÁ PRÁCE

Autor diplomové práce: Martina Juříková

Vedoucí diplomové práce: prof. Jan Štěpán

OLOMOUC 2010

Prohlašuji, že jsem diplomovou práci na téma: **Kurt Gödel a problematika umělé inteligence** zpracovala samostatně pod vedením pana prof. Jana Štěpána a uvedla jsem veškerou použitou literaturu.

V Olomouci dne 10. května 2010 .....

Martina Juříková

Děkuji panu prof. Janu Štěpánovi za odborné vedení mé diplomové práce a za cenné rady, které pro mne byly velkým přínosem. Děkuji také svým pedagogům, jejichž práce pro mne byla inspirací, děkuji za jejich pedagogické vedení a doporučení odborné literatury, zejména pak panu prof. Janu Štěpánovi, Filipu Tvrdému, Lukáši Zámečníkovi a Martině Číhalové.

OBSAH:

ÚVOD

Práce Kurta Gödela bývá považována za mezník v oblasti matematické logiky, nejvýznamnější (a zároveň nejznámější) její částí jsou Gödelovy věty o úplnosti a neúplnosti. Gödelovy věty o neúplnosti zcela změnily pohled na povahu matematiky a její možnou axiomatizovatelnost, která měla být jako taková dokázána Hilbertovým programem. Důkazem Gödelových vět byla tato snaha vyvrácena a zároveň bylo umožněno o mezích formálních systémů uvažovat zcela novým způsobem.

Gödelovy výsledky měly dopad na mnoho oblastní matematiky, logiky i filosofie a tak není překvapením, že jsou v těchto oblastech hojně využívány a interpretovány. Gödelova práce vyvolala, mimo jiné, velký ohlas a rozvoj v oblasti informatiky a umělé inteligence. Ovšem ani po téměř osmdesáti letech od Gödelova důkazu věty o neúplnosti nedošlo ke sjednocení názoru na důsledky, vyplývající z Gödelových vět, pro oblast a výzkum umělé inteligence.

V prvé řadě se zaměřím na rozhor Gödelovy práce, skrze rekonstrukci a vysvětlení jeho věty o úplnosti a obou vět o neúplnosti se budu snažit nastínit jejich vliv a důsledky pro oblast informatiky a umělé inteligence. Z toho důvodu se v další části své práce zaměřím na vazbu mezi Gödelovými teorémy a prací Alana Turinga. Zaměřím se nejen na teoretické základy a možnosti Turingových strojů, ale prostřednictvím Turingovy provokativní otázky, po možnosti myšlení strojů, se propracuji k problematice možnosti či nemožnosti simulovat myšlení mechanickými modely a následně i k otázce, jak souvisí vědomí s myšlením a inteligencí.

Následně se zaměřím na využití Gödelových vět, a jejich interpretaci, v diskuzi o možnostech AI. Jako zástupci opozice k AI budou sloužit J. R. Lucas a Roger Penrose, kteří ve své argumentaci věty o neúplnosti používají, analyzován proto bude také způsob, jak s výše uvedenými větami ve svých důkazech pracují. Pro kritické zhodnocení a reflexi jejich závěrů bude sloužit řada textů autorů, kteří se zapojili do diskuze k AI – skrze kritiku či analýzu prací zmíněných dvou autorů.

Ústřední otázkou, která se bude prolínat celým textem, přitom zůstává, zda je užití Gödelových vět pro diskuzi kolem AI relevantní a samotný způsob užití vět, v argumentaci výše uvedených autorů, korektní.

Pro svou práci budu využívat jak příspěvků, k výše naznačené problematice, které byly publikovány v českém překladu, tak většího množství textů, které do češtiny přeloženy nebyly. Domnívám se, že kvůli vysoké míře odbornosti, mnohdy zdůrazněné náročným způsobem argumentace, bude přínosnější nepřekládat, dosud nepřeložené, citované pasáže z původního jazyka. Obávám se totiž, že by mohlo dojít k dezinterpretaci významu ústředních pasáží. Přesto věřím, že text nebude touto dvojjazyčností negativně poznamenán a naopak citované pasáže přispějí k výstižnosti textu.

1. KURT GÖDEL – VĚDECKÁ PRÁCE A JEJÍ DOPAD

Kurt Gödel, jeden z nejvýznamnějších logiků 20. století, se narodil v Brně 28. dubna 1906, kde také žil až do roku 1923, kdy absolvoval německé gymnázium. V následujícím roce se odstěhoval do Vídně, kde také od roku 1924 navštěvoval tamní univerzitu, zpočátku s úmyslem zaměřit se na studium fyziky. Rozhodnutí změnit svou studijní specializaci na matematiku bylo v mnohém ovlivněno přednáškami Philippa Furtwänglera a Heinricha Gomperze, který přednášel také historii filosofie. V roce 1926 začal Gödel, v doprovodu svého profesora matematiky Hanse Hahna, navštěvovat schůzky Vídeňského kruhu. Gödel se však do debat nezapojoval, možná právě proto, že již v této době byl přesvědčeným platonistou. V roce 1929 ukončil své studium odevzdáním disertační práce, ve které definoval svůj důkaz úplnosti predikátové logiky prvního řádu. Krátký příspěvek ke svému teorému o úplnosti přednesl poté v Královci v roce 1930 na konferenci o základech matematiky, kde se také poprvé zmínil o svém důkazu neúplnosti, který – jako jediného – zaujal Johna von Neumanna.

Svou práci o neúplnosti publikoval v roce 1931 jako habilitační práci a na jejím základě se stal docentem Vídeňské univerzity v roce 1933. V akademickém roce 1933/34 navštívil Princeton, kde o svých výsledcích přednášel a mimo jiné se seznámil s Albertem Einsteinem, s kterým udržoval dlouholeté přátelství. V Princetonu se od roku 1940 zdržoval až do konce života, zpočátku jako dočasný člen Ústavu pro pokročilá studia, od roku 1946 jako člen stálý a konečně roku 1953 jako profesor. Do Evropy se už nikdy nevrátil a v roce 1948 získal americké občanství. V průběhu let pak obdržel řadu čestných akademických titulů (doktorát z Harvardské a Yalovy Univerzity) a další prestižní ocenění, např. i cenu Alberta Einsteina. Od roku 1958 přestal své závěry publikovat a nevedl dokonce už ani žádné studenty. Jeho zdravotní problémy, mnohdy vyvolané neurotickými obtížemi a paranoidním strachem z otravy jídlem, vedly k dlouhodobé podvýživě, která vyústila v jeho smrt 14. ledna roku 1978.[1]

[1]Podrobnosti z Gödelova života jsou popsány např. v Malina, Novotný (1996).

1.1. PŘEDMĚTY GÖDELOVA ZÁJMU

Kromě věty o úplnosti a dvěma větám o neúplnosti, kterým budu věnovat ve své práci samostatný prostor, se Gödel věnoval také teorii množin, která byla na počátku 20. století dominantní součástí základů matematiky a blízce s logikou souvisela. Důležité jsou hlavně jeho práce týkající se hypotézy kontinua a axiomatizace teorie množin.[2] V oblasti matematiky se Gödel dále věnoval problému nekonečněhodnotovosti intuicionistické logiky, v důsledku požadavku L. E. J. Brouwera na konstruktivní vedení důkazů v matematice.

Intuicionistická logika nepřipouští např. důkazy sporem, zákon dvojité negace, ani zákon vyloučeného třetího. Klasická výroková logika je dvouhodnotová a v závislosti na tom byly ze strany logiky intuicionismu zkoumány možnosti vícehodnotových pravdivostních tabulek. Gödel demonstroval nemožnost těchto snah a to pomocí přidání axiomu $(A \rightarrow B) \vee (B \rightarrow A)$ k intuicionistické logice. Na základě tohoto přidaného axiomu byla tato „nově vzniklá“ logika zkoumána M. Dummettem a v důsledku toho byla definována fuzzy logika.

Gödel se v průběhu svého života zabýval také filosofií matematiky. Byl přesvědčen, že matematické objekty existují nezávisle na axiomech a metodách vědy. Již od studií byl Gödel platónským realistou a zastával názor, že metody a axiomy matematické objekty nevytvářejí, pouze je popisují. Skrze toto přesvědčení věřil v nezávislá tvrzení, jež ačkoli jsou v rámci formálního systému nerozhodnutelná, jsou přesto pravdivá, nebo nepravdivá. Intuicionismus zastává názor, že pokud je tvrzení nerozhodnutelné, nemá pravdivostní hodnotu. Oproti tomu Gödel věřil, že se o pravdivostní hodnotě rozhodnout dá, i kdyby pouze skrze matematickou intuici. Tento svůj názor se snažil obhájit v tzv. Gödelově programu, ve kterém měly být rozhodnutelná všechna nerozhodnutelná tvrzení teorie množin, a to na základě řešení paradoxu množin skrze tzv. nedosažitelné kardinály.

Další částí Gödelovy práce, jež měla filosofický přesah, byla jeho rekonstrukce ontologického důkazu boží existence, na které pracoval kolem roku

[2] Základní přehled problematiky např. v Běhounek (2006).

1970, jednalo se o formalizaci Leibnizovy varianty Anselmova ontologického důkazu formalizované v modální logice S5 druhého řádu.[3]

Dalším oborem, kterému se Gödel věnoval již od dob svých studií, byla fyzika. V roce 1949 Gödel napsal článek k Einsteinovým sedmdesátým narozeninám, ve kterém pojednával mimo jiné i o možnostech cestování časem do minulosti. Možnost vycházela z Gödelova řešení Einsteinových rovnic pole v rámci obecné teorie relativity, Gödel navrhl kosmologický model, kde čas je cyklický a pohybem v prostoru by tak bylo možné dosáhnout jakékoli časoprostorové souřadnice. Einstein byl existencí Gödelova modelu znepokojen a popíral možnost jeho aplikace, protože mimo jiné byl Gödelův model v rozporu s pozorovanými vlastnostmi vesmíru.

[3] Gödelovou formalizací ontologického důkazu se zabývá např. P. Zlatoš, či P. Hájek.

1.2. GÖDELOVY VĚTY O ÚPLNOSTI A NEÚPLNOSTI

Gödelův největší vědecký přínos je spatřován v jeho větách o neúplnosti, které bývají označovány za zlomové teorémy matematické logiky. Podle Gödela bylo však jen otázkou několika měsíců, kdy by byly závěry, obsažené v jeho pracích, vyřčeny někým jiným. Počátkem 20. století byla témata jako „paradox lháře" (která pro své řešení také využívají rozlišení mezi tvrzením a tvrzením o tvrzení) velice oblíbená – zpracovávali je např. Zermelo, Tarski či Skolem. Gödel však se svými větami o neúplnosti přišel jako první a to ve velice čisté formální podobě.

Dříve však, než se zaměřím na Gödelovy věty o neúplnosti, které budou východiskem pro další části práce, je třeba krátce zmínit i jeho větu o úplnosti. Zároveň pokládám za vhodné letmo zmínit 2. problém Hilbertova programu, který je pro zřetelnost výkladu nezbytný.

1.2.1. OTÁZKA BEZESPORNOSTI ARITMETIKY

Problém důkazu bezespornosti aritmetiky byl druhým z dvaceti tří problémů, kterými se David Hilbert zabýval ve své přednášce v průběhu 2. Mezinárodního matematického kongresu v Paříži v roce 1900.[4]

Aritmetikou, v souvislosti s Hilbertovým programem, rozumíme formální systém aritmetiky reálných čísel, vybudovaný na základě axiomatického systému, kde jsou reálná čísla vymezena jako soubor objektů tvořící archimedovsky uspořádané těleso, doplněné o předpoklad nemožnosti rozšíření systému o další objekty – bez zpochybnění původních axiomů.

Bezespornost systému by byla dokázána, pokud by pomocí konečně mnoha logických úsudků nebylo možné vyvodit z axiomů daného systému důsledky, jež by

[4] Hilbertův program spočíval ve: *„1) formalizaci jazyka vlastních matematických, ale i jiných disciplín, 2) jejich převedení do axiomatického tvaru, 3) následném důkazu bezespornosti příslušných formalismů finitistickými prostředky."* Kolman (2008) s. 524. Podrobnější rozbor viz Kolman (2008) a Hilbertův druhý problém viz Bečvář (1971).

byly vzájemně kontradiktorické. Druhý problém, který byl třeba dokázat, lze tedy formulovat následovně: *„Prokázat konzistentnost axiomů aritmetiky. Systém je konzistentní, jestliže z něj nevyplývají žádné logické spory.“*[5] V této první fázi řešení tohoto problému však nebyly Hilbertem jasně vymezeny prostředky, které by bylo možné při důkazu bezespornosti použít. Bezespornost systému axiomů však představovala nutnou podmínku pro možnost přesné a úplné charakterizovatelnosti axiomů samotných. V době, kdy se Hilbert o tento důkaz pokoušel, však neexistovalo striktní odlišení obsahových a formálních stránek axiomatických teorií, stejně tak jako nebyla vymezena hranice mezi jazykem, ve kterém jsou výsledky matematiky formulovány, a metajazykem, v němž o výsledcích hovoříme.

K tomu, aby mohl D. Hilbert zkonstruovat svůj absolutní důkaz bezespornosti je nutná tzv. úplná formalizace deduktivního systému. Výrazy, které takový deduktivní systém obsahuje, musí být zbaveny veškerých významů a v důsledku toho se stát prázdnými znaky. Manipulace a kombinovatelnost těchto znaků je pak přesně stanovena pravidly systému, který může být následně označen za systém znaků zvaný „kalkul“. Jde tedy o přesně formalizovaný systém založený na transformaci teorémů z přesně definovaných postulátů. Samotné výroky tohoto systému jsou prázdné a významu nabývají až výroky provedené o tomto systému (tzv. formule). Tyto výroky však již do výše načrtnutého kalkulu nepatří, jsou součástí metamatematiky, tedy jazyka o matematice.

Ve dvacátých letech 20. století se na základě kritiky ze strany intuicionistů, zaměřené na nekonstruktivní zacházení s nekonečnými matematickými objekty, Hilbert k důkazu bezespornosti axiomatických systémů vrátil. Axiomatizace byla v důsledku striktního užití finitních prostředků nahrazena formalizací. Jak však později vyplynulo z Gödelovy 1. věty o neúplnosti, úplnost formálního systému nelze dokázat prostředky, které by mohly být plně formalizovány v rámci systému samotného. Z toho v zásadě plyne nemožnost dokázat úplnost aritmetiky, jako celé teorie. Jinými slovy můžeme říci, že Hilbertova snaha dokázat skrze bezespornost aritmetiky konzistentní povahu celé matematiky a metod, které matematika užívá,

[5] Goldsteinová (2005) s. 120.

byla – s důrazem na nutnost užití finitních prostředků vyvození, kdy za finitní prostředky pokládáme pouze ty, jež jsou formalizované v PM, později v PA – neúspěšná, ale přesto pro budoucí vývoj matematické logiky zásadní.

1.2.2. GÖDELOVA VĚTA O ÚPLNOSTI

Větu o úplnosti nějakého formálního systému lze vyjádřit jako tvrzení *všechny platné formule lze v tomto formálním systému odvodit*. Predikátový počet prvního řádu je základní logický kalkul, ve kterém platí, že všechny platné formule jazyka prvního řádu jsou, v dané axiomatice této logiky, dokazatelné. Axiomatika prvořádové logiky byla navržena Hilbertem a Ackermannem v roce 1928, kdy také formulovali problém její úplnosti, kterým se zabýval Gödel, tedy dokázat větu:

Každá tautologie klasické prvořádové logiky je dokazatelná v Hilbertově-Ackermannově systému axiomů.[6]

V NÁSLEDUJÍCÍCH KROCÍCH SE POKUSÍM GÖDELŮV DŮKAZ REKONSTRUOVAT:[7]

Tautologie je obecně platná formule – formule je tautologií, jestliže je pravdivá v každé struktuře, tedy při každé interpretaci. Příklady tautologií[8]:

1) $\varphi \to (\psi \to \varphi)$

2) $\big(\varphi \to (\psi \to X)\big) \to ((\varphi \to \psi) \to (\varphi \to X)$

3) $(\neg\varphi \to \neg\psi) \to (\psi \to \varphi)$

4) $(\forall x)\varphi(x) \to \varphi(c)$[9]

[6] Znění věty převzato z Běhounek (2006).

[7] Z velké části je rekonstrukce důkazu převzata z Malina, Novotný (1996).

[8] Pod čísly 1) – 5) myslíme schémata formulí, nikoli jednotlivé formule.

[9] C je konstanta nebo vhodná proměnná, $\varphi(c)$ znamená formuli, která vznikne z formule $\varphi(x)$dosazením c za x.

5) $(\forall x)(\nu \rightarrow \psi) \rightarrow (\nu \rightarrow (\forall x)\psi)$

Nějaké tautologie nyní můžeme, pro potřebu objasnění Gödelovy věty o úplnosti, zvolit za logické axiomy (v rámci Hilbertově-Ackermannově systému axiomů) a dále zvolme dedukční pravidla, pomocí kterých z tautologií jistého tvaru vytvoříme jiné tautologie (konkrétně pravidlo *modus ponens* a pravidlo *generalizace*)[10]

Jestliže každá formule φ_i je axiom[11] nebo φ_i bezprostředně vyplývá z některých předchozích členů posloupnosti $\varphi_1,\ldots, \varphi_{i-1}$, pak je posloupnost formulí $\varphi_1, \ldots, \varphi_n$ důkazem v predikátové logice prvního řádu.

Je-li $\varphi_1, \ldots, \varphi_n$ důkaz, pak každá formule φ_i je tautologií. Z toho plyne, že φ je dokazatelná, pokud je posledním členem nějakého důkazu. Tedy každá dokazatelná formule je zároveň tautologií.

Tím jsme dokázali obecnou platnost tautologie v rámci výše definovaného systému.

Gödelova věta o úplnosti dokazuje, že predikátový počet prvního řádu je úplný logický kalkul – to znamená, že dokazatelné jsou právě všechny tautologie. Dokazatelnost je tedy ekvivalentní pravdivosti (obecné platnosti).

Úplnost kalkulu však nemůže být zaměňována s úplností teorie. T je úplná teorie, pokud pro každou uzavřenou formuli φ T dokazuje φ nebo T dokazuje $\neg\varphi$. Jinými slovy teorie T rozhoduje každý výrok (dokazuje jeho pravdivost, nebo ho vyvrací).

PŘEDVEDEME ZESÍLENÝ DŮKAZ VĚTY O ÚPLNOSTI:

Teorie je dána množinou formulí zvaných speciální axiomy (příslušné teorie).

[10] Modus ponens je pravidlo odloučení: z formulí $(\varphi \rightarrow \psi), \varphi$ bezprostředně odvodíme ψ. Pomocí pravidla generalizace z formule φ bezprostředně odvodíme $(\forall x)\varphi$.

[11] V rámci Hilbertově-Ackermannově systému axiomů.

Struktura M je modelem teorie T, jestliže každý speciální axiom teorie T je pravdivý v M.

Důkaz v teorii T je posloupnost $\varphi_1, \ldots, \varphi_n$ formulí, jejíž každý člen je buď logický axiom, nebo je speciálním axiomem teorie T nebo vyplývá z některých předchozích – podle některého z dedukčních pravidel.

Proto je-li M modelem teorie T a formule φ je dokazatelná v T, pak φ je pravdivé v M.

SILNÁ VĚTA O ÚPLNOSTI ZNÍ:

Formule φ je dokazatelná v teorii T, právě když φ platí v každém modelu teorie T.

K důkazu dále potřebujeme definovat teorii T jako spornou:

T je sporná, jestliže je v T dokazatelná formule φ a zároveň je dokazatelná $\neg\varphi$.

LEMMA[12] O EXISTENCI MODELU (pouze pro bezesporné teorie):

Každá bezesporná teorie má model.

Pokud T nedokazuje φ, pak teorie $T \cup \neg \varphi$ je bezesporná a tedy má model M, ale neobsahující pravdivou formuli φ. Pokud tedy neobsahuje φ, tak není úplná – nemůže φ vyvodit- dokázat.

ZÁVĚREM: Predikátová logika prvního řádu je úplný logický kalkul (dokazuje každou tautologii), ale není úplnou teorií, protože nerozhoduje všechny výroky – Hilbert doufal v důkaz úplnosti aritmetiky[13] (jako celé teorie), což jak bude ukázáno dále, je vyvráceno Gödelovými větami o neúplnosti.

[12] Lemma je zde pomocným tvrzením, které nebudeme dokazovat, slouží však jako mezikrok v důkazu.

[13] *„Jistota Hilbertovy teze o poznatelnosti všeho je samozřejmě založena tím, že je triviální, neboť poznání je od počátku vždy jenom poznání lidské a jakékoli jeho božské (na člověku*

1.2.3. GÖDELOVA PRVNÍ VĚTA O NEÚPLNOSTI

Gödelova první věta o neúplnosti konstatuje neúplnost určité formální teorie, která obsahuje Peanovu aritmetiku. Gödelova věta o neúplnosti jde ale aplikovat i na jiné aritmetiky, např. na Robinsonovu aritmetiku, která oproti PA neobsahuje axiom indukce. Peanova aritmetika (PA) obsahuje tři axiomy pro funkci následníka, a to: 1) nula není následníkem žádného čísla, 2) každé nenulové číslo je následníkem nějakého čísla, 3) čísla jsou si rovna, jen když jsou si rovni jejich následníci.

Peanova aritmetika dále obsahuje:

rekurzivní definici sčítání pomocí následníka $\left(x + 0 = x,\ x + násl(y) = násl\,(x + y)\right)$

násobení pomocí sčítání $(x\,.\,0 = 0,\ x\,.násl(y) = x\,.\,y + x)$

axiom indukce pro každou prvořádnou aritmetickou vlastnost ϕ, který můžeme vyjádřit jako: má-li vlastnost ϕ nula a s každým číslem i jeho následník, pak mají tuto vlastnost všechna čísla.

Gödelovu 1. větu o neúplnosti můžeme vyjádřit následujícím způsobem:

Každá bezesporná rekurzivně axiomatizovaná[14] *teorie obsahující Peanovu aritmetiku je neúplná*[15].[16]

---

nezávislé) varianty mohou být pouze více či méně oprávněné artikulace jeho aktuální lability, tj. nezávislosti na přesvědčení konkrétního jedince, nikoli na lidstvu jako celku." Kolman (2008) s. 581.

[14] *„Teorie je rekurzivně axiomatizovaná, jestliže existuje algoritmus, který pro každou formuli rozhodne, zda φ je či není axiomem teorie T."* Novotný, Malina (1996) s. 84. Nebo jinými slovy *„Teorie je rekurzivně axiomatizovaná, má-li rekurzivní sadu axiomů a odvozovacích pravidel. Pouze takové teorie lze rozumně chápat jako „finitně" zadané, neboť pouze u nich máme konečný algoritmus, jak poznat jejich axiomy."* Běhounek (2006) s. 49. Podrobná definice rekurzivní funkce a funkce rekurzivně definované viz Gödel (1931) s. 178- 179.

[15] Teorie je neúplná, pokud existuje tvrzení, které v ní nejde dokázat, ani vyvrátit – nelze tedy toto tvrzení rozhodnout. Pokud mluvíme o teorii, která je neúplná a bezesporná

V NÁSLEDUJÍCÍCH KROCÍCH ZREKONSTRUUJI GÖDELOVU PRVNÍ VĚTU O NEÚPLNOSTI:[17]

Nechť T je teorie, obsahující Q.[18]

Nechť T je rekurzivně axiomatizovaná a nechť N je jejím modelem.[19]

Pak T je neúplná teorie.[20]

Dále je pro důkaz G. věty nutné vysvětlit Gödelovu metodu aritmetizace matematiky, což znamená, že formule (jazyka matematiky) jsou posloupnosti znaků a důkaz je posloupností formulí, přičemž Gödel přiřazuje každé formuli φ a každému důkazu d v teorii T číslo *gn* (Gödel number). Pro tuto metodu můžeme použít označení Gödelovo číslování.

Základem tohoto označení byla Gödelova snaha připsat jednotlivé číslo každému elementárnímu znaku, každé formuli a každé konečné sekvenci formulí. Elementární znaky, v Gödelově systému spojena s celými čísly, patřící do základního slovníku formalizovaného kalkulu, jsou dvojího druhu: konstanty a proměnné. Kromě znaků konstant se v PA, kde je možné vyjádřit kardinální čísla, jejich sčítání a násobení, objevují i tři druhy proměnných a z nich tvořené výroky. Jedná se o číselné proměnné, výrokové proměnné a predikátové proměnné. Základem Gödelova číslování je přiřadit jednotlivé číslo každé ze složek výroku. Přesto, že lze formule vyjádřit jako posloupnost čísel, je ujednáno, že tato číselná posloupnost je vyjádřitelná i číslem jedinečným. Tím je stanovena metoda pro úplnou aritmetizaci formálního kalkulu. Protože je každý výraz v PA spojen s nějakým jediným

---

(Běhounek (2006) uvádí korektní) jsou v ní dokazatelná pouze pravdivá tvrzení, pak existují pravdivé výroky, jež v ní nejsou dokazatelné – vyvoditelné.

[16] Znění přejato z Běhounek (2006) s. 49.

[17] Důkaz rekonstruován po vzoru Malina, Novotný (1996).

[18] Jazyk teorie T obsahuje jazyk aritmetiky a T dokazuje všechny axiomy teorie Q.

[19] To znamená, že každá formule jazyka aritmetiky dokazatelná v T je pravdivá v N.

[20] Tj. existuje výrok φ, který není rozhodnut v T (T nedokazuje ani φ ani $\neg\varphi$. Což bylo dokázáno a demonstrováno v oddílu o Gödelově větě o úplnosti.

Gödelovým číslem, nabízí se možnost hovořit o jakémsi principu zrcadlení aplikovaném právě na vztah matematiky a metamatematiky. Vzhledem k možnosti přiřazení jediného čísla výrazu v PA, může být metamatematické tvrzení o formálních výrazech chápáno právě jako tvrzení o odpovídajících číslech a jejich vzájemných relacích. Tímto posunem, tedy pomocí přiřaditelného Gödelova čísla, dochází i k aritmetizaci metamatematiky.

Dále z rekurzivní axiomatizovanosti teorie T plyne, že množina *gn* formulí dokazatelných v T je definovatelná v N jistou $\Sigma - formuli$[21], kterou označíme *Dok(x)*. Z toho plyne, že T dokazuje φ, právě když $Dok(\overline{gn(\varphi)})$ je pravdivé v N.

DEFINUJEME GÖDELOVU DIAGONÁLNÍ LEMMU:

Pro každou formuli $\varphi(x)$[22] existuje uzavřená formule ψ, taková, že $\psi \equiv (\overline{gn(\psi)})$. [23]

Nakonec aplikujeme diagonální lemmu na formuli ¬Dok(x) a tím dostaneme Gödelovu diagonální formuli, kterou označíme písmenem ν, takovou, že Q dokazuje $\nu \equiv \neg Dok\ (\overline{gn(\nu)})$.[24]

Pokud by T dokazovalo ν, pak formule $Dok(\overline{gn(\nu)})$ by byla pravdivá v N a (tato formule je $\Sigma - formule$) pak by ale T dokazovalo $\neg\nu$, protože Q dokazuje $\neg\nu \equiv Dok(\overline{gn(\nu)})$.[25] Čímž bychom došli ke sporu. Protože ale bylo dáno, že T je bezesporná, jasně z toho vyplývá, že T nedokazuje ν (samozřejmě ani $\neg\nu$), přestože

[21] Na základě $\Sigma - úplnosti\ teorie$ je definováno, že Q není úplná, ale z $\Sigma - formule$ vyplývá, že každý $\Sigma - výrok$ pravdivý v N je dokazatelný v Q.

[22] Jazyka aritmetiky s jednou volnou proměnnou.

[23] ψ říká „já mám vlastnost φ“. $gn(\psi)$ je číslo formule ψ. $\overline{gn(\psi)}$ je jméno tohoto čísla v jazyce aritmetiky a φ $(\overline{gn(\ \psi\ )})$tedy říká, že číslo $gn(\psi)$ má vlastnost φ.

[24] Tedy říká „já jsem nedokazatelná“.

[25] Z $\Sigma - formule$ vyplývá, že každý $\Sigma - výrok$ pravdivý v N je dokazatelný v Q.

je *v* pravdivé v N. T je tedy neúplná teorie, jež nedokazuje všechny formule pravdivé ve standardním modelu aritmetiky.[26]

Jinými slovy můžeme říci, že první věta o neúplnosti má formu podmínkového výroku *„jestliže je formální systém aritmetiky konzistentní, pak G je nedokazatelné“*. Tuto formu upřednostňuje Goldsteinová (2005). Pro ilustraci odcituji její krátký důkaz platnosti Gödelovy první věty o neúplnosti: *„Nechť C představuje výrok: „Formální systém aritmetiky je konzistentní“. První věta o neúplnosti nám pak říká: Jestliže C, pak je G nedokazatelné. Aritmetizací výroku „G je nedokazatelné“ je samozřejmě G. Takže první věta o neúplnosti říká, že C→G a tento závěr byl dokázán ve formálním systému aritmetiky. Takže jestliže můžeme dokázat C ve formálním systému aritmetiky, ipso facto bychom dokázali G ve formálním systému aritmetiky, neboť jsme dokázali C→G. A protože bylo dokázáno, že G je nedokazatelné ve formálním systému aritmetiky, víme, že i C je nedokazatelné v tomto systému.“* [27]

Důsledkem první věty o neúplnosti je tedy tvrzení, že každá rozumná aritmetika je nerozhodnutelná – neexistuje algoritmus, který by byl schopen rozhodnout každou formuli, tedy určit, zda je v teorii T dokazatelná, či vyvratitelná.

Na základě rekonstrukce Gödelovy první věty o neúplnosti můžeme přejít k závěrečné části této kapitoly a to sice ke Gödelově druhé větě o neúplnosti.

1.2.4. GÖDELOVA DRUHÁ VĚTA O NEÚPLNOSTI

Gödelovu druhou větu o neúplnosti můžeme vyjádřit jako:

Žádná bezesporná rekurzivně axiomatizovaná teorie T obsahující Peanovu aritmetiku nedokazuje formuli Con(T) vyjadřující její formální bezespornost.[28]

[26] Tedy pro přirozená čísla.

[27] Goldsteinová (2005) s. 158 – 159.

[28] Znění přejato z Běhounek (2006) s. 50.

Při rekonstrukci důkazu budeme stejně jako u předešlých dvou vět využívat textu Petra Hájka v Novotný, Malina (1996).

Buď T libovolná „rozumná“[29] teorie.

Uvnitř T lze vyjádřit tvrzení o bezespornosti teorie T pomocí formule *Dok*.[30]

Výrok vyjadřující bezespornost teorie označíme jako Kon (konzistence).

Gödel dokázal, že formule Kon je ekvivalentní jeho diagonální formuli ν, tedy že T dokazuje $Kon \equiv \nu$ protože T nedokazuje ν, nedokazuje ani Kon.

„Ryze syntaktické aspekty formálního systému k důkazu konzistence samy o sobě nestačí. Nepostačují ani k důkazu všech pravdivých aritmetických výroků vyjádřitelných v systému (první věta o neúplnosti), ani k poskytnutí důkazu vnitřní konzistence (druhá věta o neúplnosti).“[31]

Mnou uvedená varianta důkazu Gödelovy druhé věty o neúplnosti je zásadním způsobem vázána na důkaz (a postup důkazu), který byl rekonstruován v předchozí části kapitoly – bez rekonstrukce Gödelovy první věty o neúplnosti se tento důkaz může zdát nekompletní. Jen pro ilustraci proto naznačím i jiný způsob důkazu Gödelovy druhé věty o neúplnosti.

Například Nagel a Newman rekonstruují druhou větu (společně s první) o neúplnosti následujícím způsobem (Nagel, Newman (2006)).[32]

---

[29] Rekurzivně axiomatizovaná, obsahující Q, splňující podmínky PA, která má za model N, je nerozhodnutelná – nelze zúplnit pomocí žádného algoritmu.

[30] Tak, že formálně vyjádříme neúplnost teorie skrze nemožnost dokázat všechny formule, tedy skrze $\neg Dok(\overline{gn(\varphi)}) \vee \neg Dok(gn(\overline{\neg\varphi}))$ pro nějakou pevně uzavřenou formuli φ.

[31] Goldsteinová (2005) s. 162.

[32] Je třeba upozornit, že autoři pracují s formálním systémem PM (Rusell-Whiteheadův systém), nikoli PA jak je využíváno v tomto textu – po vzoru Hájka (Novotný, Malina (1996)). Také Gödelovo číslování vysvětlují samostatně a pro důkaz s ním již pracují jako s premisou. Protože byl v práci náležitě rozebrán formální důkaz vět (podle Hájka), nebudu na tomto místě uvádět další formální důkazy ve všech jejich krocích – interpretace by měla být dostačující.

(i) Lze zkonstruovat formuli G systému PM, jež reprezentuje matematické tvrzení „Formule G je s použitím pravidel PM nedokazatelná“. Tato formule je formálně zastoupena Gödelovým číslem, které je v systému PM nedokazatelné a je proto konstruována jako „Formule, která má Gödelovské číslo g, není dokazatelná“.

(ii) Gödel ovšem prokázal, že formule G je dokazatelná jen tehdy, pokud je dokazatelná ¬G. Formálně dokazatelné formule G a ¬G však odporují bezespornosti PM a proto pokud je systém PM bezesporný, nelze v rámci něj odvodit G ani ¬G.

(iii) Přesto, že je G formálně nerozhodnutelná formule, je pravdivou aritmetickou formulí.

(iv) Z toho, že G je pravdivá, ale v rámci PM formálně nerozhodnutelná formule vyplývá, že systém PM je neúplný.

(v) Následně lze zkonstruovat formuli A v systému PM „Systém PM je bezesporný“, která ovšem reprezentuje metamatematické tvrzení. Formule „A→G“ je v rámci PM dokazatelná, ale A není v systému PM dokazatelná. Z toho vyplývá, že bezespornost systému PM nemůže být v rámci systému dokázána (vyvozena).

Tato interpretace vět o neúplnosti pracuje explicitně s rozlišením matematického a metamatematického tvrzení, upozorňuje tedy na nutnost rozšíření tříd pravidel odvozování – má-li se stanovit bezespornost systému PM, ale samozřejmě mimo systém a nikoli v rámci systému samotného. Důkazy nelze „zrcadlit“[33] uvnitř systémů, kterých se týkají. Důkazy bezrozpornosti mimo systém tudíž nejsou finitistické – nevyhovují proto předpokladům Hilbertova programu.

[33] Nagel a Newman při svém důkazu pracují s termínem zrcadlení v rámci systému PM ve smyslu reprezentace tvrzení a to skrze Richardův paradox – není možné bezezbytku definovat všechny termíny odkazující k aritmetickým vlastnostem, neboť existují nedefinovatelné termíny, ze kterých vychází definování termínů vyvozovaných. Tomuto problému a jeho využití v interpretaci se však nebudu ve své práci věnovat.

1.2.5. DŮSLEDKY GÖDELOVÝCH VĚT

Jak už bylo naznačeno výše, jedním z nevyvratitelných důsledků Gödelových vět o neúplnosti bylo zamítnutí Hilbertova programu, tedy představ o možnostech úplné axiomatizace matematiky jako úplného, konzistentního (a svou konzistenci dokazujícího) systému a mimo jiné také možnost redukovat infinitní metody na finitní.

Ve své době očekávatelná věta o úplnosti je z historického hlediska velice přínosná, dokazuje totiž úplnost prvořádového predikátového počtu – *„Logickou platnost prvořádných formulí lze popsat konečnou mechanickou aplikací konečných pravidel, tedy naše omezené konečné prostředky kupodivu plně popisují celou nekonečnou třídu i nekonečných matematických struktur z hlediska prvořádové platnosti.“*[34]

Věty o neúplnosti a nerozhodnutelnosti jsou velice ceněným přínosem pro logiku, matematiku a další příbuzné obory, ceněny jsou však také metody, díky nimž jsou věty dokazatelné, které v mnohém obohatily jak teorii modelů, teorii rekurze, tak obzvláště metamatematiku.

Obzvláště důležité (v souvislosti s touto prací) jsou však důsledky pro oblast informatiky a teorie algoritmů. Rekurzivní funkce, která byla intuitivně užívána, ale definována až Gödelem, formálně vyjadřuje principiální algoritmickou počitatelnost. Není proto divu, že metody užité v Gödelových důkazech našly své využití ve výpočetním modelu Turingova stroje. Následující část práce se proto bude věnovat právě využití Gödelových vět v informatice a teorii algoritmů a dále se také bude věnovat interpretaci a významu vět pro oblast umělé inteligence.

[34] Běhounek (2006) s. 53.

2. GÖDELOVY VĚTY A UMĚLÁ INTELIGENCE (AI)[35]

Přesto, že jsou Gödelovy věty o neúplnosti formálně korektně dokazatelné, v některých jeho současnících přetrvával dojem, že se jedná o pouhé logické hříčky na způsob paradoxu lháře a jeho závěry byly odsunuty stranou. Např. Jacques Herbrand v roce 1931 došel k závěru, že Gödelovy věty o neúplnosti nedokazují, že je problém rozhodnutelnosti nemožný kladně vyřešit.

Profesor M. H. A. Newman se ve svých přednáškách o základech matematiky ke Gödelovým větám často vracel a v roce 1934, kdy jeho třídu navštěvoval Alan Turing, vyjádřil problém rozhodnutelnosti jako „hledání mechanického postupu pro testování platnosti určitého výroku“, který zůstal nevyřešen. Právě tato „mechaničnost testování“ přiměla Turinga k pokusu o řešení Hilbertova problému rozhodnutelnosti.

Jak uvádí Leavitt (2007), Newman popsal situaci následovně: *„Cílem Hilbertova programu v otázce rozhodnutelnosti bylo ve dvacátých a třicátých letech nalezení obecného postupu aplikovatelného na jakýkoli matematický výrok vyjádřený zcela symbolickou formou, který by dokázal rozhodnout, zda je daný výrok pravdivý, či nikoli. První ránu těmto vyhlídkám na nalezení nového kamene mudrců zasadil Gödel svou větou o neúplnosti (1931), která jasně ukázala, že v žádném dostatečně bohatém logickém systému se pravdivost nebo nepravdivost A nerovná dokazatelnosti A nebo negace A. Stále tu však zůstává možnost nalezení mechanického postupu pro rozhodnutí, zda je v daném systému formálně dokazatelné*

[35] *„Podle Marvina Minského je umělá inteligence věda, jejímž úkolem je naučit stroje, aby dělaly věci, které vyžadují inteligenci, jsou-li prováděny člověkem. … V tomto smyslu je i Minského definice umělé inteligence zaměřena příliš jednostranně. Na druhé straně ani letadla za letu nemávají křídly jako ptáci, takže často „inteligentní“ chování systému považujeme za dostačující, i když procesy probíhající v počítači nejsou totožné s procesy v lidském mozku.“* Berka (2008) s. 9. Jiný pohled na problematiku můžeme vyjádřit citací *„Již klasická logika dovedla charakterizovat intelektuální činnosti, kterými se vyznačují myslící bytosti. Těmito činnostmi mají být podle klasiků analýza, syntéza, indukce, dedukce a analogie. Zmocnit se podstaty těchto činností matematickými prostředky, aby bylo možno jejich provádění svěřit počítačům, to je problémová oblast, která bývá nazývána umělou inteligencí.“* Lukasová (1995) s. 1.

A, nebo negace A, nebo ani jedno z nich. Mnozí byli přesvědčeni, že žádný takový postup není možný, ale až Turing se rozhodl tuto nemožnost demonstrovat přesně a názorně.“[36]

2.1. ALAN TURING A TURINGOVY STROJE

Alan Turing (údajně – viz Leavitt (2007)) získal svou první představu Turingova stroje na základě významů slova „mechanický“, jež původně znamenalo manuální činnost prováděnou lidmi a ve třicátých letech asociovalo „stroj“, zároveň slovo „počítač“ bylo významově spojeno s osobou, jež provádí výpočty – používající algoritmy[37]. K práci výpočtáře bylo ve třicátých letech užíváno pomůcek, jako počítadel či sčítacích strojů, jež byly však čistě pasivní. Žádné výpočetní stroje tedy neexistovaly.[38]

Turing své výsledky poprvé prezentoval v roce 1937 v článku *On computable numbers, with an application to the Entcheidungsproblem*[39] v časopise *Proceedings of the London Matehmatical Society*. Ústřední myšlenkou je otázka, jakými možnými způsoby lze vypočítat nějaké číslo – Turing ve své argumentaci definuje počitatelná čísla jako reálná čísla, jejichž vyjádření v desítkové soustavě jsou počitatelná

[36] Leavitt (2007) s. 48.

[37] Algoritmus je postup pro řešení určitého druhu problémů, který je prováděn pomocí konečného počtu přesně definovaných kroků. *„Při rozhodování splnitelnosti, event. logické platnosti formule A se často hovoří o rozhodovacích algoritmech. Obecně se jimi rozumí procedura rozhodování, zdali určitý objekt je prvkem určité množiny objektů.“* Lukasová (2003) s. 32.

[38] V devatenáctém století se pokoušel o sestrojení parou poháněného „analytického stroje“ Charles Babbage, který však ve svém snažení nebyl úspěšný. Má se za to, že Babbage nedošel k řešení vkládání instrukcí ve stejném matematickém jazyce, v jakém měl být prováděn výpočetní postup těchto instrukcí.

[39] *O vyčíslitelnosti s ohledem na problém rozhodnutelnosti*.

konečnými prostředky a číslo je počitatelné, jestliže je jeho vyjádření možné zapsat pomocí stroje.[40]

Počítací stroj v této počáteční fázi Turingovy práce je označován jako „a-machine“, což bude podstatné i pro pozdější interpretace. V zásadě je však popsán stejně, jako „univerzální“ Turingův stroj[41], sice hlava – spojená s konečnou řídící jednotkou, která čte a případně i přepisuje pásku se symboly, jež je oboustranně nekonečná. Stroj, který dokáže generovat počitatelnou posloupnost je definován jako necyklický, a to oproti cyklickému stroji, který negeneruje žádnou smysluplnou posloupnost. Důležitá otázka je[42], zda je možné navrhnout takový stroj, který by analyzoval jiný stroj a rozhodl, zda je či není cyklický – přes tuto otázku se Turing dostává k otázce, zda existuje stroj E, „E“ jako „Entcheidungsproblem“, který dokáže určit, zda je daný výrok dokazatelný, či nikoli. Skrze spletitý důkaz dochází Turing k závěru, že není možné, aby takový stroj existoval – problém rozhodnutelnosti tedy nelze vyřešit skrze „mechanický“ výpočet – neexistuje tedy obecná metoda, kterou by bylo možno problém rozhodnout.

---

[40] *„The „computable“ numbers may be described briefly as the real numbers whose expressions as a decimal are calculable by finite means. … Acording to my definition, a number is computable if its decimal can be written down by a machine.“* Turing (1937) s. 230.

[41] Turingův stroj, zkráceně TS, je definován jako šestice $M = (Q,\Sigma, \Gamma, \delta, q0, F)$, kde:Q je konečná neprázdná množina stavů, Σ je konečná neprázdná množina vstupních symbolů, Γ je konečná neprázdná množina páskových symbolů, kde $\Sigma \subseteq \Gamma$ a v $\Gamma - \Sigma$ je (přinejmenším) speciální znak (prázdný znak [blank]), $q0 \in Q$ je počáteční stav, $F \subseteq Q$ je množina koncových stavů, $\delta : (Q - F) \times \Gamma \rightarrow Q \times \Gamma \times \{-1, 0, +1\}$ je přechodová funkce.

[42] Na kterou navazuje problém zastavení (halting problem), který můžeme vyjádřit otázkou: *„je možné rozhodnout o tom, zda výpočet kteréhokoli programu skončí v konečném čase, nebo nikoli?“ … „Tento zajímavý problém dostal i své vlastní jméno: halting problem neboli problém zastavení (míněno: zastavení Turingova stroje). Bohužel se však ukázalo, že nemá řešení. Přesněji že není algoritmicky rozhodnutelné, zda se libovolný Turingův stroj (alias počítač s libovolným programem) v konečném čase zastaví, či nikoli. Tomuto výsledku je ovšem třeba správně rozumět. Říká totiž, že nikdy nebude existovat algoritmus, který by problém zastavení Turingova stroje řešil pro libovolný program (resp. pro libovolný Turingův stroj). Tím ovšem není vyloučeno, aby takovýto algoritmus existoval pro určitou podmnožinu všech možných programů (Turingových strojů) - neexistuje pouze takový, který by byl jeden a fungoval spolehlivě pro jakýkoli program“* Peterka (1994).

Přibližně ve stejné době, kdy se Turing pomocí „Turingovy teze“ vyrovnával s problémem rozhodnutelnosti, byl vydán článek, ve kterém stejný problém, ale jinou metodou řešil Alfonso Church. Newman nakonec Turingovi doporučil článek přesto publikovat a s prof. Churchem vyjednal jejich společnou práci na problému rozhodnutelnosti na Princetonské univerzitě.[43]

Na základě podobné argumentace byla teze teorie vyčíslitelnosti pojmenována po Churchovi i Turingovi jako Church-Turingova teze[44], která ve stručnosti říká, že jakýkoli vypočitatelný problém (tedy každý možný výpočet) je možné vyjádřit pomocí algoritmu Turingova stroje. Algoritmus je tedy v běžném chápání s Turingovým strojem ekvivalentní.

2.1.1. MOHOU STROJE MYSLET?

Ústřední problém, ze kterého vychází mnohé interpretace Gödelových vět ve vztahu k AI, je obsažen v Turingově práci *Computing Machinery and Intelligence* publikované v roce 1950. Ještě předtím však představím méně známý[45] Turingův text *Intelligent Machinery*, jehož ústřední téma je spojováno s otázkou, „zda mohou být

[43] Respektive Turing pod vedením Churche sepisoval svou disertační práci a také se zabývali Gödelovými větami o neúplnosti, které se úzce vázaly jak na Churchovu, tak Turingovu práci. Church však odmítal Gödelovy výsledky, možná i proto, že Gödel neocenil práci Churche a v roce 1946 v dopise upřednostnil práci Turinga.

[44] *„ Ke každému algoritmu je možné zkonstruovat s ním ekvivalentní Turingův stroj (s rozumným kódováním vstupů a výstupů řetězci v určité abecedě); ekvivalenci zde rozumíme podmínku, že algoritmus i Turingův stroj vydají pro tytéž vstupy tytéž výstupy.“* Jančar (2007) s. 187.

[45] Text je přístupný z Turingova digitálního archívu, kde ovšem není uvedeno, zda byl text vůbec publikován a pokud ano, tak v kterém roce a kde. Předpokládám, že byl napsán mezi roky 1938 a 1940 a to vzhledem k literatuře, ze které Turing vycházel a dále z poznámky u námitky č. 3 *„The very limited charakter of the machinery which has been used until recent times (e.g. up to 1940).“* Tamtéž s. 1.

zkonstruovány stroje, které by jevily známky inteligentního chování.“[46] Turing hned v úvodu předkládá možné námitky, proč by stroje nemohly vykazovat známky inteligentního chování.[47]

1)Neochota připustit, že by lidé mohli mít nějaké soupeře, co se týká rozumové schopnosti.

2)Náboženská víra, že jakýkoli pokus zkonstruovat takové stroje je něco jako prométheovská opovážlivost.

3)Velká omezenost dosud používaných strojů, která podporuje domněnku, že stroje jsou nutně omezeny na vykonávání jednoduchých a opakujících se úkonů.

4)Jakýkoli stroj nebude v některých případech schopen vůbec odpovědět, zatímco stále se rozvíjející lidská inteligence se zdá být schopná najít metody pro řešení problémů, které transcendují metody dostupné strojům.[48]

5)Pokud může stroj jevit známky inteligence, není to nic jiného než odraz inteligence tvůrce.

Možné námitky jsou ihned následovány Turingovými odpověďmi na tyto námitky, což považuji za velice přínosné. Stručně zrekapituluji, že námitka 1) a 2) jsou podle Turinga spojené pouze s osobním přesvědčením – není je proto třeba vyvracet. Námitka 3) je rychle vyvrácena existencí strojů ENIAC či ACE, které

[46] *„I propose to investigate the question as to whether it is possible for machinery to show intelligent behaviour. It is usually assumed without argument that it is not possible.“* Turing (1940) s. 1.

[47] Vzhledem k obsáhlosti námitek nebude citován všechen původní text, ale pouze nejrelevantnější jeho část, ale vše je citováno dle Turing (1940) s. 1 – 2.

[48] *„Recently the theorem of Gödel and related results (Godel 1, Church 1, Turing 1) have shown that if one tries to use machines for such purposes as determing the truth or falsity of mathematical theorems and one is not willing to tolerate an occasional wrong reset, then any given machine will be in some cases unable to give an answer at all. On the other hand the human intelligence seems to be able to find methods of ever increasing power for dealing with such problems „transcending“ the methods available to machines.“* Turing (1940) s. 2.

mohou provádět obrovský počet operací, aniž by se opakovaly, pokud nedojde k poruše.

Námitka 4), která se nejvíce týká našeho tématu, je Turingem vyvrácena na základě jeho přesvědčení, že neomylnost není nutnou podmínkou inteligence – což jak uvidíme v další části práce je svým způsobem protiargumentem Lucasově námitce obsažené v jeho textu, jež byl publikován až o minimálně třináct let později.[49]

Poslední námitka 5) je vyvrácena tvrzením totožným s názorem, že za objevy žáka je třeba vděčit učiteli. V tom případě by měl být učitel spokojen s úspěchem svých vyučovacích metod, ale neměl by si dělat nároky na výsledky samotné, pokud je svému žákovi přímo nepředal. Jako příklad takového stroje Turing uvádí stroj (program) hrající šachy. Tato odpověď na poslední námitku také souvisí se zbytkem textu, ve kterém Turing apeluje na metody učení strojů, které jsou podle jeho názoru klíčem k budoucímu uznání statustu inteligentního stroje.[50]

Vrátím se tedy k Turingově nejcitovanějšímu textu *Computing Machinery and Intelligence*, ve kterém je podrobně nastíněna „imitační hra“, jež je pro Turinga zástupným vysvětlením za termíny „stroj“ a „myslet“, protože uvádí, že význam těchto termínů je určen jejich běžným užitím v jazyce, význam by tedy mohl být definován až na základě nějakého statistického průzkumu, což by bylo absurdní. V důsledku uvedené imitační hry se tomuto testu „inteligence“ začalo říkat Turingův test – na základě toho, jak je stroj schopen přesvědčivě předstírat roli člověka, je mu přiřknut statut projevu inteligentního chování.

[49] *„The argument from Godel´s and other theorems rests Essentials on the condition that the machine must not make mistakes. But this is not a reguirement for inteligence.“* Odpověď na námitku pokračuje argumentací procesu učení u dětí školního věku – text je však špatně čitelný, takže je možné vydedukovat obsah, ale přepis by byl nehodnotný.

[50] Na základě této části budeme podrobněji rekonstruovat Turingovu imitační hru – rozpracovanou v Computing Machinery and Intelligence, takže imitační hru uvedenou v Intelligent Machinery necháme pro tuto chvíli stranou.

2.1.2. TURINGOVA IMITAČNÍ HRA

„Hrají tři lidé, muž (A), žena (B) a moderátor (C), který může být jakéhokoli pohlaví. Moderátor zůstává v místnosti oddělené od ostatních dvou hráčů[51]*. Úkolem moderátora je určit, kdo z obou hráčů je muž a kdo žena. Zná je pod označením X a Y a na konci hry řekne buď „X je A a Y je B“, nebo „X je B a Y je A“. Moderátor se může ptát A a B takto: C: Mohlo by mi X sdělit délku svých vlasů? Předpokládejme nyní, že X je A, takže A musí odpovědět. Úkolem A je mást C tak, aby provedl chybnou identifikaci. Může proto odpovědět: „Mám dlouhé blond vlasy.“ Úkolem třetího hráče (B) je napovídat moderátorovi. Pravděpodobně nejlepší strategií pro ženu je odpovídat správně. Může ke své odpovědi dodat: „Jsem žena, nevěř mu!“*[52]*, ale ani to mnoho neznamená, protože muž může říct něco podobného. Nyní se můžeme zeptat, co se stane, když stroj převezme úlohu A v této hře? Rozhodne se pak moderátor stejně tak často špatně, jako kdyby hru hrál muž a žena? Tyto otázky nahradily původní otázku: „Mohou stroje myslet?“*[53]

Turing svou imitační hrou „ztotožnil“ otázku „mohou stroje myslet?“ s otázkou „může stroj uspět v imitační hře?“ – dle jeho názoru jsou otázky ekvivalentní (význam jejich sdělení je identický), což vyvolalo nejrůznější otázky a následně i interpretační problémy. Slovo „myšlení“ bylo (více méně stále je) primárně spojováno pouze s živými organismy, pokud by však bylo přijato splnění imitační funkce jako kritérium myšlení, bylo by možné rozšířit význam slova „myšlení“ i na stroje a to po vzoru rozšíření užití slova „létat“ z ptáků na letadla. Problém také nastal při interpretaci uvedené pasáže „co se stane, když stroj převezme

[51] Komunikace probíhá přes prostředníka nebo psanou formou, aby moderátorovi nepomáhaly k identifikaci hlasové charakteristiky.

[52] Na tuto imitační hru nelze aplikovat řešení lhářova paradoxu – není dáno, že A pouze lže a B říká pouze pravdu.

[53] Turing (1950) s. 433 – 434. Uvedená pasáž je natolik známá, že zde nebudu citovat originální text.

roli A?“ Interpretace tohoto problému nechám stranou, protože už nejsou pro mé původní téma tolik relevantní.[54]

Skrze Turingovy úvahy se do filosofie vrací otázky po vztahu mysli a těla a po povaze myšlení obecně. Nejvyhraněněji odmítl závěry Turingovy práce John Searl, představitel biologismu, jež svůj protiargument představil v podobě myšlenkového experimentu čínského pokoje, na kterém demonstruje odlišnost mezi manipulací se symboly a porozuměním významu – jasně tedy upozorňuje na dvě úrovně jazyka – syntax a sémantiku – kdy sémantika zůstane počítači vždy nepřístupna. Dalším kritikem „rozšíření“ pojmu myšlení na stroje byl Ned Block, který upozorňoval na nutnost zvážit nejen výsledek „imitační hry“, ale také způsob, jakým bylo požadovaného výsledku dosaženo. Do třetice je za kritika považován i Roger Penrose, jehož práci však budu věnovat samostatný prostor.[55] Za protistranu v této diskuzi můžeme zmínit Daniela Dennetta a jeho teorii postojů, jež ve stručnosti říká, že člověk zaujímá určitý projektový postoj (k ostatním lidem, či věcem), na základě kterého předpokládá jejich chování. Otázku vědomí a myšlení poté problematizuje skrze determinismus a fyzikalismus.

Na základě Turingovy imitační hry je od roku 1990 udělována Loebnerova cena v oblasti AI právě za „splnění“ imitační funkce programu.[56] Je však třeba upozornit, že zatím žádný z programů nesplnil požadavky této imitační hry, pokud nebylo přesně vymezeno konverzační téma a prostředky, jakými může tazatel program „testovat“. Je však nutné upozornit, že byla mnohdy kritizována právě tato

[54] Respektive se ani nechci zabývat různými interpretacemi propojující Turingovu homosexuální orientaci s analogiemi určení pohlaví v imitační hře, které pokládám za naprosto nepřesvědčivé a zbytečné – podrobněji např. Leavitt (2007), jež Turingovu sexuální orientaci tematizuje v celé knize a do svého výkladu ji zapojuje mnohdy, dle mého názoru, až neadekvátně.

[55] Lépe řečeno jeho argumentům, ve kterých využívá Gödelových vět pro odlišení algoritmického a nealgoritmického myšlení, či projevu chování.

[56] Více o podmínkách, vítězných programech i proměnách soutěže je možné zjistit na oficiálních stránkách http://www.loebner.net/Prizef/loebner-prize.html.

metoda, kterou by se mělo o „inteligenci“ strojů rozhodovat, protože jejím hlavním kritériem je schopnost „rozumné“ verbální komunikace[57].

Náčrt problematiky kolem Turingových strojů ukončím návratem ke Gödelovým větám o neúplnosti. Turinga ke zkonstruování prvního počítače inspiroval problém rozhodnutelnosti. Stejně jako Gödel dokázal, že žádný dosti silný formální systém nemůže být úplný a dokázat svou vlastní bezespornost[58], Turing dokázal, že nelze sestrojit výpočetní stroj, jenž by problém rozhodnutelnosti dokázal univerzálně vyřešit. Turing Gödelův teorém o neúplnosti v podstatě formalizoval pomocí Turingova stroje. Byl si však velice dobře vědom námitky, která se s nemožností tohoto univerzálního stroje, řešícího problém rozhodnutelnosti, přímo pojí. Je jí matematická námitka, která upozorňuje na „omezenost“ či limitovanost strojů, právě skrze využití Gödelovy věty o neúplnosti.[59] „*Jsou určité věci, které stroje nedokážou, ale stejně tak lidé. Ovšem pokud stroj podá špatný výsledek, vzbudí to v lidech pocit nadřazenosti. Je tento pocit iluzorní? Není pochyb o tom, že je oprávněný, ale nemyslím si, že by mu měla být přikládána velká důležitost. My sami také často dáváme špatné odpovědi, což si omlouváme tím, že s velkým potěšením sledujeme podobnou omylnost na straně strojů. Nadřazenost navíc můžeme cítit i v případech, když získáme nepatrnou převahu nad jedním strojem. Nepřichází ovšem v úvahu, že bychom zvítězili nad všemi stroji najednou.*“[60][61]

---

[57] *„Nejzávažnější námitkou proti Turingovu testu je, že měří schopnost počítače simulovat myšlení, ale nic nevypovídá o vlastní inteligenci počítače. Kromě výsledného efektu by se měla posuzovat i inteligence způsobu řešení úlohy. Inteligence je tímto testem redukována na pouhou schopnost přesvědčivě lhát.“* Berka (2008) s. 9.

[58] *„The best known of these results is known as Gödel´s theorem, and shows that in any sufficiently powerful logical system statements can be formulated which can neither be proved nor disproved within the system, unless possibly the system itself is inconsistent.“* Turing (1950) s. 444.

[59] Stejně jako podobné závěry prací Churche, Kleena, Rossera a Turinga samotného.

[60] Všimněme si, že téma „nadvlády“ bude problematizováno jak u Lucase, tak Penrose.

[61] *„The result in question refers to a type of machine which is Essentials a digital computer with an infinite capacity. It states that there are certain things that such a machine cannot*

Turingova odpověď na možnou matematickou námitku může, ale nemusí být považována za přesvědčivou a vyčerpávající. Pokud bychom měli přijmout Turingovu odpověď, zdá se, že Gödelovy věty nejsou pro diskuze kolem možnosti AI relevantní. Znamená neúplnost formálního systému, či nerozhodnutelnost všech formulí, nemožnost inteligentního chování tohoto systému? Nebo v přeneseném smyslu, je nutnou podmínkou inteligence zodpovědět všechny možné otázky a úkoly? – pokud ano, jaké odpovědi budou při posuzování dostačující – pouze správné odpovědi, nebo naopak pouze správné znamená příliš přesné a tudíž neomylné, tedy opět výhradně „nelidské"? Na druhou stranu se můžeme domnívat, že právě nesplnitelný problém rozhodnutelnosti a důkaz vlastní bezespornosti a úplnosti je pro limitovanost strojů a jejich inteligenci relevantní.

Mimo jiné z důvodu nejednoznačnosti názorů na tento problém je diskuze kolem možnosti využití Gödelových vět v diskuzi o AI možná a smysluplná. V další části práce se proto zaměřím na argumenty, které Gödelových vět využívají, otázkou však zůstává, zda je těchto matematických vět využito korektním způsobem, či zda slouží pouze jako zastřešující prvek pro jinak nesmyslné a nekorektní výtky, které nemají s AI mnoho společného.

---

do. If it is rigged up to give answers to questions, as in the imitation game. There will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply. There may, of course, be many such questions, and questions which cannot be answered by one machine may be satisfactorily answered by another. We are of course supposing for the present that the questions are of kind to which an answer „Yes" or „No" is appropriate, either than questions such as „What do you think of Picasso?" … This is the mathematical result: it is argued that it proves a disability of machines to which the human intellect is not subject. The short answer to this argument is that although it is established that there are limitations stated, without any sort of proof, that no such limitations apply to the human intellect. But I do not think this view can be dismissed quite so lightly … There would be no question of triumphing simultaneously over all machines." Turing (1950) s 444- 445.

3. VYUŽITÍ GÖDELOVÝCH VĚT V DISKUZI O AI

Gödelovy věty byly v diskuzi o možnostech a omezeních umělé inteligence využity v práci J. R. Lucase a Rogera Penrose. V obou případech šlo o snahu, pomocí těchto vět, ostře vymezit hranice, které nemůže AI překročit. Podobný záměr obou autorů měl však zcela jiné motivace a jak shrnu v závěru práce, jasně se odlišují i v chápání konceptu AI, proti kterému se vymezují. Práce obou autorů vyvolala bouřlivý ohlas, pokusím se tedy naznačit i nejsilnější námitky, které mohou prozradit směr, kterým se vydává současný výzkum v oblasti AI. Domnívám se, že práce těchto dvou autorů doplněna o množství komentářů a námitek, je pro postihnutí problematiky dostačující. Je možné, že se Gödelovy věty, jako argument pro vyvrácení či podporu AI, vyskytují i u jiných autorů, ti však v mé práci zmíněni nebudou – důvodem může být špatná dostupnost jejich práce, nezapojení se do rozsáhlejších diskuzí a nebo, což je časté, jejich práce ve velké míře obsahuje opakující se myšlenky, které budou zmíněny u jiného z mnou uvedených autorů či kritiků.

3.1. VYVRÁCENÍ MECHANICISMU J. R. LUCASEM

V roce 1961 publikoval J. R. Lucas svou stať *Minds, Machines and Gödel,* ve které využil Gödelovy věty o neúplnosti k vyvrácení mechanicismu a k důkazu, že mysl nemůže být vysvětlena na základě fungování stroje. Od roku 1961 byl text několikrát přetištěn. Obzvláště přínosné je pro diskuzi k problematice vydání časopisu *Etica & Politica* z roku 2003. V tomto čísle se kromě původního Lucasova textu objevilo mnoho příspěvků reagujících na Lucasův text, doplněných o odpověď J.R. Lucase.

J. R. Lucas ve svém textu formuluje a následně obhajuje základní cíl své práce: dokázat, že lidská mysl a stroje jsou zásadně odlišné a že není možné vysvětlit lidské chování za pomoci mechanického modelu či analýzou jakéhokoli neživého systému. Předem je nutno podotknout, že text J. R. Lucase není v zásadě proti myšlence či možnosti umělé inteligence. Nesnaží se Gödelův teorém o neúplnosti využívat pro vyvrácení samotné možnosti umělé inteligence, zabývá se něčím

naprosto jiným – věty o neúplnosti užívá pouze jako prostředku pro odlišení lidské mysli od jakéhokoli formálního systému, kterým by snad mohla být neadekvátně lidská mysl popisována či vykládána. To také souvisí s jím používaným termínem „mechanicismus“, který můžeme definovat jako snahu vysvětlit lidskou mysl na základě mechanického modelu – v první řadě Lucas tedy neproblematizuje otázku, zda mohou stroje myslet, ani se nevyjadřuje k Turingově kritériu rozšíření pojmu „myšlení“ na stroje, ale svou snahu směřuje pouze na odlišení lidské mysli a jejího výkladu, od popisných mechanických modelů.

Gödelův teorém o neúplnosti musí být, podle Lucase, aplikován na stroje prvořadě, protože podstatou „být strojem“ je konkretizace, či zpředmětnění, formálního systému. Lucas ve svém textu uvádí, že jednoznačnost Gödelova teorému dokazuje, že jakýkoli formální systém je neúplný a formule G je dokazatelná pouze mimo daný formální systém. A právě to je základem pro odlišení myslící bytosti a neživé věci či stroje – možnost vystoupit mimo daný formální systém a rozhodnout o platnosti formule G, která je v rámci systému nedokazatelná. Tím je autorem obhájen názor, že mechanické modely (stroje) založené na naprogramovaném, tedy předem definovaném formálním systému, nemohou adekvátně zobrazit či napodobit lidskou mysl, protože *kybernetické stroje* jsou založeny na předen definovaném souboru pravidel, podle kterých jednají a s ohledem na okolnosti volí mezi možnostmi, které jsou předem dány. Stroje jsou tedy oproti mysli naprosto determinované.

Autor uvádí několik příkladů, na kterých demonstruje, že stroje nejsou schopny Gödelův teorém o neúplnosti překonat, nesnaží se však dokázat, že by lidská mysl byla nadřazena naprogramovanému stroji ve smyslu dokonalosti či výkonnosti. Právě naopak. Možnost chyby dělá člověka méně přesným, než je stroj, ale zároveň to opět dokládá nenapodobitelnost lidské mysli, protože žádný stroj by „vědomě“ neporušil soubor daných pravidel – to je dáno jeho prvotním naprogramováním – není schopen provést úkon, který by vedl k jeho vlastní nekonzistenci. Za to nejdůležitější autor považuje rozdílnost mysli a stroje, ne jejich vzájemné soupeření o nadvládu. Gödelův důkaz je z toho důvodu symbolickou Achillovou patou formálních systémů, není však vyloučeno, že teorém o neúplnosti

bude přístroj schopen přidat mezi axiomy a na základě toho vystavět meta-systém, se kterým bude moci formuli v rámci nového systému dokázat. Tak ostatně, podle Lucase, pracuje s Gödelovým důkazem lidská mysl. V čem je tedy rozdíl, mezi myslí a strojem či mechanickým modelem mysli? Program či přístroj, který dokáže prokázat neprokazatelnou formuli vně původního systému je novým programem, obohaceným o další formální systém obsahující teorém o neúplnosti. Lidská mysl zůstává stejnou, ať přemýšlí ve více úrovních (tedy o sobě) či nikoli.[62]

Základní myšlenkou tedy zůstává nutnost odlišení stroje a lidské mysli, z jakého důvodu? Není to proto, že by odmítal možnosti, které stroje člověku poskytují, ani proto, že by nevěřil, že v budoucnu budou moci existovat stroje natolik složité, že by se u nich mohla definovat schopnost jakéhosi základního myšlení. Lucasova motivace spočívá ve snaze zachovat či zachránit lidskou jedinečnost, svobodu vůle, a odmítnout tak determinismus.

Pokud by nebyla vyvrácena možnost simulovat lidské myšlení pomocí mechanických modelů, znamenalo by to, že lidská mysl není jedinečná, nenapodobitelná a v zásadě nenaprogramovatelná. Lucas se domnívá, že „záchrana" svobodné vůle je s vědeckým pokrokem téměř neslučitelná. Možnost simulovat lidskou mysl by se rovnala ztrátě víry ve svobodu vůle. Ačkoli připouští možnost

[62] *„We could construct a machine with the usual operations, and in addition an operation of going through the Gödel procedure, and then producing the conclusion of that procedure as being true; and then repeating the procedure, and so on, as often as required. This would correspond to having a system with an additional rule of inference which allowed one to add, as a theorem, the Gödelian formula of the rest of the formal system, and then the Gödelian formula of this new, strengthened formal system, and so on. It would be tantamount to adding. to the original formal system an infinite sequence of axioms, each the Gödelian formula of the system hitherto obtained. Yet even so, the matter is not settled: for the machine with a Gödelizing operator, as we might call it, is a different machine from the machines without such an operator; and, although the machine with the operator would be able to do those things in which the machines without the operator were outclassed by a mind, yet we might expect a mind, faced with a machine that possessed a Gödelizing operator, to take this into account, and out-Gödel the new machine, Gödelizing operator and all. This has, in fact, proved to be the case. Even if we adjoin to a formal system the infinite set of axioms consisting of the successive Gödelian formula, the resulting system is still incomplete, and contains a formula which cannot be proved-in-the-system, although a rational being can, standing outside the system, see that it is true."* Lucas (2003) s. 8.

vědeckého pokroku, zejména v oblasti zkoumání lidského mozku, domnívá se, že díky nepřekonatelnosti Gödelova teorému nebude nikdy možné, aby mechanické modely byly schopny mysl simulovat, vždy ji mohou maximálně popisovat.

3.1.1. MECHANICISNUS VS. MENTALISMUS

J. R. Lucas prostřednictvím Gödelových vět o neúplnosti vyvrací mechanicismus, a to na základě tvrzení, že program, který by byl schopen vyjádřit konzistentnost jiného programu (nižšího programu – ve smyslu toho, že tento nižší program není schopen vyjádřit svou vlastní bezespornost), není již tím původním programem. Původní program, obsahující určité výchozí axiomy, ke kterému by byl dodán Gödelův teorém (ať už vyjádřený jakkoli), by podle Lucase nebyl totožným programem, ale programem jiným – na základě přesvědčení, že jedině lidská mysl je schopna určité sebereflexe, či odstupu (nadneseně můžeme říci, že je schopna vystoupit z programu či nad systém, ve kterém je definována). S tímto názorem v zásadě polemizuje I. J. Good ve svém článku *Human and Machine Logic*. K mému překvapení však nevznáší zcela zásadní námitku a to, proč je takový program nutně pokládán za program nový – zcela odlišný od původního programu bez Gödelova teorému.

Good do jisté míry rekonstruuje Lucasův argument a dodává, že do formálního systému F (kde F=(R,A), kde A je konečná množina axiomů a R jsou pravidla usuzování, prostřednictvím kterých z daných teorémů mohou být odvozeny teorémy nové) může být dodán teorém G (mezi axiomy A), jenž je pravdivý, ale v F nedokazatelný, pokud je F bezesporný systém. Pokud je F_1 bezesporný, může být G_1 (jako pravdivé tvrzení) dodáno do množiny axiomů A_1 a nový systém F_2 bude stále bezesporný. Toto obohacování o formule G_n může pokračovat do nekonečna, kdy vždy dostaneme vyšší systém F_{n+1}. Good argumentuje, že takto „obohacovaný“ program bude vždy schopen přesáhnout možnosti lidského výpočtáře a tudíž problém mechanického modelu lidské mysli nespočívá v principiální nemožnosti takového modelu, ale v neschopnosti formálně popsat všechny vlastnosti, které by měly být do modelu mysli zahrnuty. Pokud je má interpretace správná, obávám se, že námitka

vznesená ze strany Gooda nemůže být brána za relevantní k Lucasově textu. Jak ostatně podotýká Lucas ve své odpovědi na tuto námitku. V první řadě mu nejde o soupeření lidské mysli a výpočetního programu, ale zdůraznění jejich odlišností a nemožnosti ekvivalence.[63]Druhou možností je, že se Good snažil svou námitkou obhájit mechanicismus a možnost úplně popsat mysl postupným obohacováním systému o jednotlivé *Gn* formule – podotýká totiž, že tyto *Gn* formule mohou být považovány za vlastnosti lidského myšlení – je však otázkou, zda měl Good opravdu namysli postupným přidáváním formulí dojít až ke komplexitě systému, či nikoli.

Na jedné straně se Lucas snaží o vyvrácení mechanicismu, který Good obhajuje, ale zásadní je způsob, jakým je mechanicismus vyvrácen, tedy s využitím Gödelových vět, které Good ve své námitce také využívá, ale zcela kontroverzním způsobem - domnívám se proto, že námitka míří špatným směrem a snaží se pouze vyvrátit Lucasův „mentalismus“ a nikoli použití Gödelových vět v Lucasově argumentaci a bohužel ani neupozorňuje na nutnost přesněji definovat odlišnost programu nižší úrovně a programu, jehož množina axiomů je obohacena o Gödelův teorém a (což Lucas nepopírá) je program schopen po tomto začlenění s teorémem pracovat.

Lucasův argument v zásadě pracuje s pojetím stroje jako čistě deterministického formalizovaného systému, tedy uzavřeného systému, jež není schopen pracovat s přidanými teorémy mezi původní axiomy. Takový stroj je roven původnímu Turingově *a-machine*, jež byl však navržen pro řešení problému

[63] *„For if mechanism is true, a complete specification of the mental mechanism of each human being can in principle be given. But once given, it proves inadequate, in that it cannot produce as true the Gödelian formula G, which a human being can see to be true. This is no `hollow victory' for the human being, since this was the specification the mechanist put forward. It is of course true that a second computer, that is, another computer, could do as well as a human operator on this test: but that is not relevant, for it was not that computer that was supposed [156] to match the human being. And if the mechanist now makes it relevant, by shifting his ground and saying that it is the second computer that matches the human being, then there is another Gödelian formula which the second computer cannot produce as true but which a human being can. In my original paper I took the argument as far as omega. Good takes a further step to omega + 1, …”* Lucas (2003) s. 42.

rozhodnutelnosti – vyčíslitelnosti. Kromě této základní verze *a-machine* však nejen Turing rozlišuje jiné Turingovy stroje např. *c-machine*, který je oproti *a-machine* definován jako stroj s částečnou determinací. Částečná determinace stroje, s využitím axiomatických systémů, je závislá na volbě externího operátora (jež není podrobněji definován) – stroj tedy není úplně determinován a může se uvažovat o začlenění Gödelových formulí mezi axiomy, které mimo jiné *a-machine* nemůže z axiomů odvodit. Hlavní otázkou tedy zůstává, jak moc je determinován stroj v Lucasově argumentaci a zda je vyloučeno působení programu samotného na sebe – tedy pracovat na dvou úrovních – využívat přidaných Gödelových formulí mezi axiomy.

Například McDermott tvrdí, že digitální počítače nejsou v pravém slova smyslu formálními systémy, na které se vztahují Gödelovy věty – požadovaná konzistence pro ně není určující. Digitální počítače jsou formálním systémem, ale liší se od formálního systému, ke kterému se vztahují jeho výpočty. Můžeme tedy McDermottův přístup interpretovat jako tvrzení o formálním systému počítače, který pracuje s formálním systémem – svých výpočtů, tedy na dvou různých úrovních, kde v rámci vyššího formálního systému jsou prováděny výroky o výpočtech nižšího formálního systému. Bezrozpornost je tedy dokazována o formálním systému, ale v zásadě jiným formálním systémem.[64]

[64] *„Digital computers are formal systems, but the formal systems they are almost always distinct from the formal (or informal) systems that their computations relate to. To analyze a digital computer as a formal system is merely to express its laws of operations in the form of transition rules among discrete states. When we take the inputs and outputs of the computer to refer to various real states of affairs, then it need not be the case that there exists a consistent or sound formal system C such that whenever the computer concludes Q from P, the conclusion is licensed by C. Nothing prevents me from writing a program that, given any input P, prints out "P and not-P." There is, of course, a formal system S that stipulates that exactly this event is to occur, but this formal system is not about the entities mentioned in sentence P. If it's about anything, it's about the states of the computer, and nothing more. To make this point vivid, note that S, even though it prints out self-contradictory sentences, is "consistent," considered as a formal system, because it never says that the computer is to be in two distinct states at the same time. Consistency is essential to a formal system, because in almost all formal logics anything at all follows from a contradiction. Consistency is not, however, essential to computers."* McDermott (1995) s. 5 – 6.

Další možný přístup, k vyvrácení Lucasova argumentu proti mechanicismu, je předložen Davidem Lewisem v textu *Lucas against Mechanism*. Lucas tvrdí, že jeho výstupy o úplnosti aritmetiky nemohou být nikdy kompletně duplikovány a dokázány strojem, Lewis na základě tohoto tvrzení podrobuje Lucase stejné zkoušce, ve které má dokázat, že je schopen dokázat úplnost a bezespornost „Lucasovy" aritmetiky – má tudíž dokázat, že je schopen vyvodit a ověřit všechny platné teorémy své aritmetiky – musí tedy disponovat schopností dokázat tato tvrzení na obecné úrovni. Pro uznání korektnosti jeho závěrů bychom, podle Lewise, museli být obeznámeni s Lucasovou metodou verifikace a stejně, i kdyby byly Lucasovy závěry platné, nemusely by být nutně platné pro všechny „stroje". Lewisův protiargument obdobně analyzuje i Havlík (2007), který dochází k závěru, že jeho argument i přes svou nepříliš silnou přesvědčivost můžeme považovat za velice originální – domnívám se, že nejzajímavějším krokem Lewisovy argumentace je podrobení „zkoušce" samotného Lucase v rámci jeho, tedy přeneseně lidské, „aritmetiky". Nutno však podotknout, že Gödelův důkaz je právě takovým teorémem, který člověk může vyvodit a dokázat.[65] Pro přijetí Lewisova stanoviska bychom však museli být ochotni připustit ekvivalenci formálního systému programu a „systému" nějakého člověka.

Relevantností Gödelova důkazu v Lucasově argumentaci se dále zabývá David Coder ve svém příspěvku *Gödel's Theorem and Mechanism.* Podle Codora Gödelův teorém dokazuje pouze to, že ne všechny mysli mohou být vysvětleny mechanickým modelem, ale Gödelův důkaz nelze brát jako dostačující podmínku obecné nemožnosti lidské mysli vysvětlit pomocí mechanických modelů, protože

[65] *„Although Lucas has good reason to believe that all theorems of Lucas arithmetic are true, it does not yet follow that his potential output is the whole of Lucas arithmetic. He can produce as true any sentence which ha can somehow verify to be a theorem of Lucas arithmetic. If there are theorems of Lucas arithmetic that Lucas cannot verify to be such, then, his potential output falls short of Lucas arithmetic. For all we know, it might be the potential output of a suitable machine. To complete his argument that he is no machine - at least, as I have restated the argument - Lucas must convince us that he has the necessary general ability to verify theoremhood in Lucas arithmetic. If he has that remarkable ability, the he can beat the steam drill - and no wonder. But we are given no reason to think that he does have it."* Lewis (2003) s. 2.

hypoteticky mohou existovat lidské mysli, které budou schopny, podobně jako stroje, dedukovat teorémy z axiomů formálního systému, ale nebudou schopny vyvodit Gödelův důkaz o neúplnosti.[66] Jen pro upřesnění je vhodné podotknout, že Codor sám nezastává pozici mechanicismu, o to více je přínosná jeho výtka, kterou činí vůči Lucasově vymezení mechanicismu. Mechanicismus je v Coderově pojetí matematickým konceptem, kde chování stroje je zcela určeno způsobem, jak je zkonstruován – nutnou podmínkou tedy je, že konstrukce stroje určuje jeho operace, které jsou algoritmické a algoritmus je především mechanická procedura, toto vymezení však neznamená, jak podle Codera vysvětluje Lucas, že postačující podmínkou pro vymezení stroje je schopnost provádět mechanické procedury.[67]

Posledním Lucasovým kritikem, o kterém se na tomto místě zmíním, je David L. Boyer, který kritizuje nejen Lucasův výklad mechanicismu, ale především postup, jakým je Lucasem, dle Boyera, neoprávněně vyvrácen. Ve svém textu *J. R. Lucas, Kurt Gödel, and Fred Aster* Boyer uvádí, že Lucas užívá naprosto neoprávněně a vágně termíny jako „důkaz", „dokazatelný v S", „vyvození", „výpočet", „vědět" a mnoho dalších, kterých navíc užívá pro neadekvátní výklad a následné vyvrácení mechanicismu. Lucas má údajně za cíl dokázat, že důkaz proveden strojem nebude nikdy ekvivalentní důkazu provedenému člověkem. Podobně jako již zmíněný Lewis i Boyer upozorňuje na nejasnost důkazu vlastní

[66] *„But it is not true that Goedel's theorem proves this. At most, Goedel's theorem proves that not all minds can be explained as machines. Since this is so, Goedel's theorem cannot be expected to throw much light on why minds are different from machines. Lucas overestimates the importance of Goedel's theorem for the topic of mechanism, … One man is smart enough to exercise some ingenuity in the deduction of theorems in number theory. Given a simple theorem, he may see how to prove it. But he is not smart enough to follow Goedel's proof."* Coder (2003) s. 1.

[67] *„It seems to me that Lucas misunderstands "mechanical" in this way. He thinks of being unable to calculate except according to mechanical procedures as a sufficient condition for being a machine. But either this leaves out of account the fact that one important part of something's being a machine is that its construction determines its operation. … Our idea of a machine is just this, that its behaviour is completely determined by the way it is made and the incoming "stimuli": there is no possibility of its acting on its own: given a certain form of construction and a certain input of information, the nit must act in a certain specific way. We, however, shall be concerned not with what a machine must do, but with what it can do."* Coder (2003) s. 3.

konzistence Lucasem, ve stejném smyslu, jako má dokázat svou bezespornost a úplnost počítač. Mimo to však zdůrazňuje metodu, kterou by měl tento důkaz být proveden. Principielně, pokud by chtěl Lucas porovnávat oba důkazy, musela by existovat jednotná forma pro oba důkazy. Zásadní je však Boyerovo přesvědčení, že Lucasova obava z mechanicismu je zbytečná – schopnosti počítače a výpočetních modelů totiž, údajně, nesouvisí s možnou ztrátou svobody vůle, lidského sebeuvědomění a podobně. Mechanicismus se tedy nesnaží člověka nahradit a zpochybnit jeho hodnoty. Ohledně využití Gödelových vět zaujímá Boyer jednoznačně kritický postoj, nejsou pro vyvrácení mechanicismu (a v rozšířeném smyslu pro určení inteligence počítačů)ani překážkou, ani prostředkem a způsob, jakým jich využívá Lucas je dle Boyera bezpředmětný – slouží pouze pro důkaz nemožnosti dokázat programem svou vlastní bezespornost, zatímco člověk je tohoto důkazu schopen – Boyer však dodává, že důkaz Gödelových vět o neúplnosti není jak pro mechanicismus, tak pro fungování strojů zásadní.[68]

3.1.2. ZÁVĚRY PLYNOUCÍ Z DISKUZÍ NAD LUCASOVÝM TEXTEM

Z diskuzí nad Lucasovým textem vyplývá, že jak z pozice mechanicismu, tak mentalismu je vznášen požadavek na preciznější vymezení pojmu „stroj“. Lucas ve svém textu tento pojem dostatečně nedefinuje, naopak pracuje s intuitivním pojetím tohoto pojmu a zdůrazňuje pouze definitivnost stroje – ve smyslu konečnosti stroje v prostoru i čase, dále ve smyslu jeho určitelnosti (ve smyslu účelu) a konečně zdůrazňuje jeho definitivnost ve smyslu deterministického chování stroje, které je určené souborem pravidel.

Dále můžeme vyvodit i nutnost jasněji definovat jak mechanicismus, tak mentalismus, protože pro přehlednost závěrů, a k nim vedoucím argumentům, zdá se nestačí vědět, že oba názory či konstrukty stojí vůči sobě v opozici.

[68] *„Whatever may be involved in deciding whether mechanism is true, two points should be clear. First, the mechanist owes us a terrific definitional debt, which must be paid before the question can finally be settled. Second, Gödel´s theorem owes us no grounds to bet one way or the other on the final outcome.“* Boyer (1983) s. 158.

Mentalismus, naznačený pouze tím, že mysl je v zásadě jedinečná a neuchopitelná díky přítomnosti tajuplné entity, která jako taková zůstává fyzice utajena, je stejnou měrou pro diskuzi zavádějící jako mechanicismus, ztotožnitelný se snahou vysvětlit podstatu mysli jako mechanický proces. Na základě textů kritiků se mohu domnívat, že Lucas užívá tento termín záměrně neobjasněn, aby skrze jeho „umělé“ vyvrácení mohl obhajovat jedinečnost a nepopsatelnost lidské mysli.

Z mého pohledu nejdůležitější součást Lucasovy argumentace, tedy využití Gödelových vět o neúplnosti, se zdá být problematickou oblastí stejnou měrou, jako naznačený problém definovatelnosti pojmu „stroj“ a rozlišení mezi mechanicismem a mentalismem. V zásadě je však autory textů užití Gödelových vět v Lucasově argumentaci kritizováno. Kritizován je i formálně nedokonalý, mnohdy zavádějící, důkaz obsahující Gödelovy věty, sloužící k vyvrácení mechanicismu. Zásadní však není Lucasův důkaz samotný, (který můžeme, ale také nemusíme, přijmout, protože Lucas ve svém článku užívá velice zjednodušenou formulaci Gödelovy věty o neúplnosti, která může být zavádějící, jak uvádí např. Putnam (1995)), ale otázka, zda je využití Gödelových vět relevantní. Z pohledu kritiků Lucasova textu se zdá být jejich využití neopodstatněné, což můžeme přisoudit i Lucasovi, protože jeho důkaz se pohybuje v kruhu – chce dokázat jedinečnost a odlišnost lidské mysli na základě předpokladu, že lidská mysl je jedinečná a odlišná od mechanického modelu.

K otázce po relevantnosti užití Gödelových vět se vrátím na samém konci mé práce, předtím se však budu věnovat dalšímu autorovi, jehož využití Gödelových teorémů v diskuzi o možnostech AI vzbudilo bouřlivý ohlas. Závěrečné shrnutí se bude týkat obou autorů, takže nepokládám za vhodné, aby se shrnutí na různých místech opakovala.

3.2. ROGER PENROSE – „TO NEVYPOČITATELNÉ V NAŠEM VĚDOMÍ“

Jak uvádí Roger Penrose v knize *Třetí kultura* Johna Brockmana, vždy ke všemu přistupoval vědecky, věřil, že je třeba uchopit a pochopit proces myšlení vědeckým pojmoslovím. Z důvodu, že ve vědě, kterou používáme dnes, není pro vědomé jevy žádné místo či vysvětlení, je třeba založit vědu novou, lépe řečeno přepracovat vědu současnou tak, aby v ní bylo pro vědomé projevy místo. Současná věda je, podle Penrose, vystavěna na představě, že co nelze simulovat pomocí počítače, to není pro vědu předmětem. Mnohé činnosti mozku lze simulovat pomocí počítačů, ale oproti těmto simulacím funguje vědomí odlišně – ne v tom smyslu, že by přesahovalo fyziku, ale ve smyslu toho, že přesahuje fyziku tak, jak ji chápeme dnes. Vědomí je v zásadě nevypočitatelné povahy – nemá tedy povahu výpočtu.

Důvodem, proč Penrose věří, že vědomí zahrnuje nevypočitatelné „ingredience“ jsou Gödelovy věty o neúplnosti. Penrose shrnuje podstatu Gödelových vět následujícím způsobem: *„… který na příkladě Gödelova teorému názorně ukázal, že způsob, jímž dokazujete formální nedokazatelnost jistého výroku, zároveň odhaluje jeho pravdivost. Do té doby jsem o Gödelově větě měl jen matné povědomí: říká, že lze vytvořit tvrzení, která nemůžete dokázat, ať už použijete jakýkoli systém předem stanovených pravidel. Najednou mi bylo jasné, že pokud věříte v pravidla, která používáte, pak rovněž musíte věřit v pravdivost tohoto tvrzení, i když důkaz jeho pravdivosti je mimo dosah samotných pravidel.“*[69] Nevypočitatelná povaha vědomí není v zásadě něčím náhodným či nepochopitelným, jen se nedá vyjádřit pomocí výpočtu či algoritmu. Penrose svou myšlenku demonstruje připomenutím Hilbertova 10. problému, týkajícího se řešení algebraických rovnic v oblasti celých čísel, který nemá obecné algoritmizovatelné řešení (podobně jako problém rozhodnutelnosti). K samotné formulaci svého názoru o nevypočitatelnosti vědomí byl Penrose vyprovokován až názory Marvina Minského, kterými přesně, to však Penrose neuvádí.

Penrose se také vymezuje proti možnosti vysvětlit nevypočitatelné procesy mysli prostřednictvím kvantových počítačů, které údajně ani nevypočitatelné operace

[69] Brockman (2008) s. 239.

neprovádějí, jakkoli jsou kvantové procesy pro fungování mozku důležité, k nevypočitatelným událostem dochází až na přechodu mezi kvantovou a klasickou úrovní, tento přechod je však, podle Penrose, mimo chápání současné kvantové mechaniky. Penrose se proto snaží vystavět novou fyziku, ve které bude mít hlavní postavení teorie twistorů. V této „nové fyzice“ bude možné definovat nespočetnou teorii kvantového měření, prostřednictvím kterého by snad bylo možné vysvětlit i vědomí. Vzhledem ke vzdálenosti tohoto tématu a tématu mé práce se však nebudu podrobněji zabývat ani Penrosovou kritikou současného pojetí vědy, ani jeho představou vědy nové. Zaměřím se pouze na Penrosovo využití Gödelových vět, prostřednicvím kterých se snaží dokázat nevypočitatelnou povahu lidského vědomí a nemožnost simulovat lidské vědomí pomocí počítačů.

3.2.1. VĚDOMÍ, VÝPOČET A CHÁPÁNÍ MATEMATIKY

Roger Penrose ve svém chápání matematiky vychází ze tří předpokladů 1) fyzikální svět je zcela popsatelný, v principu, matematikou – z toho vyplývá, že fyzikální svět se matematikou řídí, 2) ve fyzickém světě se nenacházejí duchovní objekty, které by na fyzický svět neměly vazbu, 3) našim duševním schopnostem je v nějakém smyslu dostupný každá prvek platonského světa. Pro vysvětlení vazby mezi fyzikálním a duševním světem je pro Penrose nezbytné objasnit pojem „vědomí“ – který jako takový může být v zásadě vysvětlen vědeckou metodou, ale není možné, aby „vědomí“ bylo ve všech svých aspektech simulováno počítačem.

Pasivní projevy vědomí (uvědomování si) zahrnují vnímání barev nebo harmonie, užívání paměti a podobně. Oproti pasivním projevům existují aktivní projevy vědomí, mezi něž patří svobodná vůle a jednání na základě svobodné vůle. Podstatným rysem vědomí, který však Penrose vyčleňuje jak z pasivních, tak aktivních projevů, je porozumění (či vhled). Porozumění je důležité pro objasnění pojmu „inteligence“. Penrose nedefinuje žádný z používaných termínů, domnívá se však, že pro pochopení a vyložení souvislostí mezi těmito termíny nejsou definice pojmů zásadní.

Ke vztahu mezi vědomým myšlením a výpočtem je, podle Penrose, možno zaujmout jedno z uvedených stanovisek A) – D)[70]. Výpočtem Penrose rozumí: *„Výpočet je to, co dělá počítač počítačem. Reálný počítač je limitován omezenou pamětí, já budu ale hovořit o idealizovaném počítači zvaném Turingův stroj, který se od normálního počítače liší jen tím, že má nekonečně velkou paměť a může počítat nekonečně dlouho, aniž udělá chybu nebo přestane fungova., ..., Výpočet nemusí sestávat jen z aritmetických operací, mohou v něm být obsaženy i logické operace.“*[71]

A) Veškeré myšlení je jen výpočet. Pocit vědomí je vyvolán čistě provedením příslušného výpočtu. Jedná se o stanovisko silné AI či výpočetního funkcionalismu.

B) Uvědomování si je rysem fyzikální aktivity mozku. Avšak zatímco sama fyzikální aktivita se dá simulovat početně, počítačová simulace pocit vědomí nevyvolává. Podle Penrose je stanovisko B propagováno J. R. Searlem – mozková aktivita je simulovatelná, ale pocit uvědomění je vázán na fyzikální konstrukci mozku.

C) Příslušná fyzikální aktivita vyvolává pocit vědomí, ale tuto aktivitu nelze plně simulovat výpočetně. Stanovisko zastávané Penrosem – ve fyzikální aktivitě mozku je něco, co leží za hranicemi vypočitatelnosti. Stanovisko C) se dá dále dělit na silné (současná fyzika nestačí k popisu procesu uvědomování si) a slabé (k nalezení nevypočitatelného nemusíme za hranice současné fyziky).

D) Vědomí se nedá vysvětlit pomocí fyzikálních, informatických nebo jiných vědeckých pojmů. Stanovisko říká, že je chyba se pokoušet o výklad vědomí pomocí vědeckých pojmů – není to možné.

Penrosovo členění přístupů ke vztahu výpočtu a vědomí na tomto místě uvádím, více méně pro úplnost, v následující části práce bude podrobně probrána Penrosova argumentace – s využitím Gödelových vět, které jak uvidíme, mají podpořit stanovisko C), které také zastává Penrose.

[70] Znění stanovisek A) – D) převzato z Penrose (1999) s. 88. Stejné znění je uvedeno i v Penrose (1994) s. 12.

[71] Penrose (1999) s. 92.

3.2.2. GÖDELOVY VĚTY A PROBLÉM ZASTAVENÍ

Gödelovy teorémy využívá Penrose při řešení problému zastavení (Penrose sice nepoužívá termínu „problém zastavení“, ale z jeho postupu („důkazu“) je patrné, že se o tento problém jedná). Pro svůj důkaz Penrose zavádí algoritmus A, o kterém je přesvědčen, že ve skutečnosti neexistuje, jenž je schopen ověřit platnost π_1-vět (π_1-věta je tvrzení, že výpočet nikdy neskončí). Ověření proběhne tak, že se A zastaví – vydá výstup o platnosti π_1-věty.[72] *„Výpočty, které působí na určité číslo n, lze v podstatě chápat jako počítačové programy. Můžete udělat seznam počítačových programů a každému z nich přiřadit číslo, řekněme p. Takže do svého univerzálního počítače vložíte nějaké číslo p, počítač se rozběhne a provede p-tý výpočet pro jakékoli číslo n, které jste zvolili.“* [73]

Již z počátku důkazu můžeme být překvapeni, jakým způsobem pracuje Penrose s Gödelovým teorémem: *„Důležitým rysem těchto typů výpočtů je právě to, že závisí na přirozeném čísle n. To je totiž ústředním bodem argumentu známého jako Gödelův. Podám ho v základní podobě, kterou mu dal Alan Turing, použiji jej však trochu jiným způsobem než on.“*[74]

Pokračujme v důkazu (problému zastavení) tak, jak činí Penrose:[75]

Pokud A(p,n) skončí, pak $C_p(n)$ neskončí.

---

[72] *„Suppose, then, that we have some computational procedure A which, when it terminates, provides us with a demonstration that a computation such as C(n) actually does not ever stop. We are going to try to imagine that A encapsulates all the procedures available to human mathematicians for convincingly demonstrating that computations do not stop. Accordingly, if in any particular case A itself ever comes to an end, this would provide us with a demonstration that the particular computation that it never to does not ever stop.“* Penrose (1994) s. 72.

[73] Penrose (1999) s. 95.

[74] Penrose (1999) s. 95.

[75] Důkaz je ve stejném znění uveden jak v Penrose (1999) s. 96 – 97, tak v Penrose (1994) s. 74 – 77, ale s tím rozdílem, že v Penrose (1994) je použito q místo p.

Předpokládejme nyní, že položíme p = n. Což je Cantorův diagonální postup. Dojdeme tedy k závěru: Pokud A(n,n) se zastaví, pak $C_p(n)$ nezastaví.

A(n,n) je tedy funkcí jednoho čísla a musí tedy být obsaženo mezi $C_p(n)$. Můžeme tudíž zavést program k, jenž je identický s A(n,n).

Poté platí: $A(n,n) = C_k(n)$.

Položíme tedy číslo n = k a dojdeme k $A(k,k) = C_k(k)$. (Druhá část Cantorova diagonálního postupu).

Z původního: „Pokud A(p,n) skončí, pak Cp(n) neskončí“ učiníme závěr:

Pokud se A(k,k) zastaví, $C_k(k)$ se nezastaví.

A(k,k) je však totéž jako $C_k(k)$. Došli bychom tedy k logickému sporu: pokud se $C_k(k)$ zastaví, pak se nezastaví.

Penrose pokračuje: *„My jsme však předpokládali, že procedura A je taková, že v některých případech se A(p,n) nezastaví a Cp(n) se také nezastaví. To tedy musí nastat v případě A(k,k) = Ck(k). Protože se však výpočet A nezastavil, nevíme, zda se Ck(k) zastaví. Určitá výpočetní procedura tedy nemůže plně zahrnout veškeré matematické uvažování vedoucí k rozhodnutí, zda se určité výpočty nezastaví, tedy k stanovení pravdivosti π_1-vět. To je jádro Gödelova-Turingova argumentu v té formě, v jaké jej budu potřebovat.“*[76]

Penrose tedy pouze přepisuje problém zastavení, založený na desátém problému Hilbertova programu. Pokud tomu tak není, domnívám se pak, že Penrose ve svém důkazu naprosto neadekvátně slučuje Gödelovu větu, Turingovu tezi a Cantorův „důkaz“ nespočetnosti množin, zároveň ani jedno z uvedeného nepodává korektně a v souvislostech, ve kterých byly důkazy vystavěny.

Gödel se, dle mého názoru, ve svých důkazech nezabýval platností π_1-vět, ale důkazem neúplnosti formálního systému (existuje tvrzení, které není v rámci

[76] Penrose (1999) s. 97.

systému dokazatelné, ani vyvratitelné) a nemožnosti dokázat bezrozpornost systému v rámci systému samotného.

Turing ve své tezi zase řešil problém rozhodnutelnosti, což s Penrosovým důkazem souvisí nejtěsněji, ale už Turing dokázal, že žádný algoritmus, který by byl schopen obecně řešit problém rozhodnutelnosti, nemůžeme definovat a i na základě toho je definována Church-Turingova teze, jenž říká, že každý VYPOČITATELNÝ problém může být algoritmizován – vyjádřen algoritmem. Problém zastavení je proto algoritmicky nerozhodnutelný.

Cantorův „důkaz" nespočetnosti množiny reálných čísel, k níž je použito diagonální metody, která se používá při řešení problému zastavení, v zásadě říká, že každá podmnožina čísel je buďto spočetná (existuje její očíslování přirozenými čísly), nebo má mohutnost kontinua (což znamená existenci jejího zobrazení na množinu reálných čísel). Obecný důkaz však nestanovil.

Abych nebyla neoprávněně kritická, je důležité vzít v potaz i možnost, že Penrose skutečně používá pro svou argumentaci problém zastavení, v takovém případě je však nutné podotknout, že tohoto problému si byli vědomi již Turing, Church a další, když ukázali algoritmickou nerozhodnutelnost některých úloh, či problémů. Tak je například algoritmicky nerozhodnutelná i predikátová logika prvního řádu, u níž byla dokázána její úplnost jako formálního systému – *„Řečeno jazykem teorie rekurze, množina všech tautologií je sice rekurzivně spočetná, není však rekurzivní.*" [77] Algoritmická nerozhodnutelnost problému zastavení je v důsledku negativním řešením desátého Hilbertova programu, (který žádal mechanickou metodu rozpoznání řešitelnosti diofantických rovnic), v důsledku tedy říká, že neexistuje program, který by pro každý program a jeho vstupní data rozpoznal, zda se zastaví či zacyklí. Penrose však pokračuje: *„Můžeme se ptát po jeho opravdové síle. Jasně z něho vyplývá skutečnost, že matematický vhled nemůže být zakódován do nějakého výpočtu, o kterém bychom si byli jisti,*

[77] Běhounek (2006) s. 54.

že je správný. Tento závěr bývá sice někdy zpochybňován, mně se však zdá nezvratný.“[78]

Penrose připojuje na podporu svého závěru i několik citací Turinga i Gödela, neuvádí však, odkud citace pocházejí a vyvozuje z nich závěry typu: *„Myslím, že se Turing domníval, že lidská mysl užívá algoritmů, ale tyto algoritmy jsou prostě chybné, tedy ve své podstatě nejisté. Takový přístup se mi moc nelíbí, protože v tuto chvíli se nezabýváme otázkou, jak člověk získává inspiraci, nýbrž problémem, jak může sledovat určitou argumentaci a rozumět ji.“*[79]Vzhledem k tomu, že Penrose neuvádí zdroj citovaných pasáží a jejich obsah zdá se mi být v mnohém rozporný oproti textům, které jsem pro svou práci používala sama, nebudu závěry, které z nich Penrose vyvozuje brát za relevantní. A to i z důvodu, že tvrzení, která Penrose pronáší např. o Gödelově přesvědčení (ve věci AI) např.: *„Podívejme se, co říkal Gödel. Ten by b mém schématu byl osobu D. Vidíme, že třebaže oba, Turing i Gödel, vycházeli ze stejného matematického důkazu, došli k zcela opačným závěrům, co se týče jeho obecných důsledků. Nicméně ačkoli Gödel ve skutečnosti nevěřil, že matematický vhled se dá redukovat na nějaký výpočet, neuměl tuto možnost rigorózně vyloučit.“*[80] jsou mnohdy v rozporu nejen s primární literaturou, ze které jsem vycházela, ale i s literaturou sekundární, která uvádí citace v korektním znění.

Vraťme se však k Penrosově argumentaci. I když připustíme, tu nejrozumnější možnost, že Penrose využívá problému zastavení, k čemu tento důkaz nemožnosti algoritmicky rozhodnout problém zastavení vlastně používá? Penrose tvrdí, že: *„Gödelův argument se týká určitých speciálních výroků o číslech. Co nám Gödel říká, je, že žádný systém početních pravidel nemůže plně charakterizovat všechny vlastnosti přirozených čísel. Přestože taková pravidla neexistují, každé dítě ví, co to přirozená čísla jsou. … Dítěti nemusíte dávat sadu početních pravidel – jen se snažíte, aby „pochopilo“, co to přirozená čísla*

[78] Penrose (1999) s. 97.

[79] Penrose (1999) s. 97.

[80] Penrose (1999) s. 98.

jsou."[81] Penrose dodává, že člověk, díky schopnosti „být si vědom věcí" má přístup do platonského světa, kde jsou potřebné znalosti k pochopení již obsaženy. Matematické porozumění tedy není obecně výpočetní věc, ale je založeno na naší, lidské, schopnosti či vlastnosti vědomí.

Nevýpočetní charakter (čehokoli) Penrose obhajuje pomocí nevypočitatelného modelu vesmíru na hraní, na kterém dokládá, že ač je čistě deterministický, je nevypočitatelný, protože ho není možné simulovat (podle důkazu Roberta Bergera týkajícího se nemožnosti počítačově rozhodnout zda nějaká polyominová množina vydláždí rovinu v tomto modelu). Domnívám se, že to však není hodnověrné vysvětlení, protože nevýpočetní charakter, který má být dokázán, je zároveň předpokladem, stejně tak nevýpočetní charakter má sloužit jako vysvětlení nemožnosti simulovat lidské vědomí, přičemž za předpoklad nevypočítatelnosti je pokládána nemožnost něco simulovat. Rysu nevypočitatelnosti Penrose dále využívá pro naznačení a nutnosti založení nové fyziky, která by mohla náležitě vysvětlit lidské vědomí, které má být v zásadě nevypočitatelné.

Co můžeme dále považovat za důležité je Penrosovo rozlišení vypočitatelnosti a determinismu, které Penrose využívá k naznačení problematiky kolem svobody vůle, množství témat, které ve spojení s tímto složitým tématem Penrose naznačuje, ale ve skutečnosti neřeší, je natolik obsáhlé, že není v možnostech mé práce je adekvátně analyzovat. Vzhledem k tomu, že při své argumentaci již Gödelových vět na tomto místě neužívá, nechám tuto oblast jeho práce stranou.

[81] Penrose (1999) s. 99 – 100.

3.2.3. KOREKTNOST A RELEVANTNOST PENROSOVÝCH ARGUMENTŮ

Penrosovy příspěvky k problematice umělé inteligence vzbudily bouřlivý ohlas, v mnohém ještě větší ohlas, doprovázen přísnou kritikou, než tomu bylo v případě J. R. Lucase a jeho textu *Minds, Machines and Gödel.* Vzhledem k podrobnějšímu zpracování mých námitek, se na tomto místě nebudu zmiňovat o všech významných komentátorech Penrosových tvrzeních, pokusím se však vystihnout to nejpodstatnější. Nebudu na tomto místě reflektovat kritiky vztahující se k Penrosově pojetí nové fyziky či kritice kvantové teorie. Nebudu se zde ani zabývat ohlasy, které vyvolala Penrosova představa systému mikrotubulů a cytoskeletu, jako možného umístění vědomí a podobně. Zaměřím se tedy pouze na kritiku užití Gödelových teorémů, či přesněji, jak Penrose užívá, Gödelovy-Turingovy teze ve svém pojetí nevypočitatelnosti vědomí.

Zajímavým příspěvkem ke korektnosti užití Gödelových vět je text *Penrose´s Gödelian argument* Solomona Fefermana, který uvádí, že samotný Gödelův teorém je užit korektním způsobem, ale následná práce s ním má jisté formální nedostatky, které skrze mnohá zobecnění znevažují výsledný argument – například Penrosova dvojí notace konzistentnosti systému[82] a dále záměna Gödelova teorému za Rosserův teorém,[83] dále uvádí nejeden případ, kdy Penrose nesprávně formule vyvozuje.[84]

---

[82] *„ In Penrose´s account of Gödel's incompleteness theorem, he says (p. 91) that if a formal system is sound then "it is certainly ω -consistent". This is a different notion of soundness from that on pp. 74–75, since ω-consistency is stronger than consistency, i.e. than soundness for Π1 sentences. Penrose does not explain here what is meant by this new notion of soundness, but implicit in what he says is soundness for all (arithmetical) sentences [cf. the discussion of p. 112below]."* Feferman (1995) s. 7.

[83] *"Penrose further says here that he will use the notation 'G(F)' for the [formal] assertion that F is consistent. He then says that Rosser's theorem tells us that if F is consistent then G(F) is not a theorem of F; but that is what Gödel's 2nd incompleteness theorem tells us, not Rosser's. Penrose further muddies the picture by saying that he will "not bother to draw a clear line between consistency and ω-consistency" in most of his discussions, but that "the version of the G¨odel theorem that I [Penrose] have actually presented in sec. 2.5 is essentially the one that asserts that if F is ω-consistent, then it cannot be complete, being unable to assert Ω(F) as a theorem.""* Feferman (1995) s. 7.

Feferman sám popírá komputacionismus a zastává pozici platonika ve filosofii matematiky, ale přesto upozorňuje, že Penrose ze svého platonistického přesvědčení nejenže udělal premisu ve svém důkazu, ale považuje ho i za všeobecně přijaté názorové stanovisko, jež je součástí jeho důkazu.

Pravděpodobně nejcitovanějším textem, vyjadřujícím nesouhlas, je práce Drew McDermotta *Penrose is Wrong*, ve které McDermott v zásadě nesouhlasí s celým Penrosovým konceptem a způsobem, jakým se snaží vyvrátit možnosti AI a potencionální schopnosti AI vysvětlit a „nasimulovat" mysl a vědomí. Pro téma mé práce je však rozhodující, jakým způsobem McDermott popírá relevantnost využití Gödelova důkazu v této diskuzi, či přinejmenším způsob, jakým tak činí Penrose, tedy k důkazu nevypočitatelnosti vědomí.[85]

Podle McDermotta Penrose tvrdí, že matematici nepoužívají (z obecného hlediska) algoritmus pro ověření a důkaz pravdivosti tvrzení,[86]nýbrž jsou schopni „nahlédnout" pravdivosti důkazu právě díky matematickému vhledu, který je záležitostí lidského vědomí – jeho nevypočitatelné ingredienci. Musíme však upozornit, že pravdivost a dokazatelnost často takto dokazujeme u problémů, u nichž předpokládáme bezespornost dokazovaného systému. Z toho důvodu Penrose, aby byl problém náležitě obecný, uvažuje o matematicích jako skupině, která se na pravdivosti systému shodne – ne na pravdivosti formálního systému, jako je např. Peanova aritmetika, ale na pravdivosti matematických tvrzení o něčem. McDermott uvádí příklad toho, že ne vždy může být „obecné" stanovisko o nějakém problému

---

[84] *„Penrose says here that if F* and F** are obtained from F by adjoining G(F) and ¬G(F) resp. as axioms, and if F is consistent then F* and F** are both consistent. This is correct for F** by ordinary logic, but not for F*."* Feferman (1995) s. 8.

[85] *„Penrose stakes everything on his analysis of Gödel's Theorem. This analysis is all wrong, but what's striking is how much he tries to hang on it. Penrose assumes that there is a single attribute called "consciousness" that accounts for insight, awareness, and free will. Hence, if he can show that computers lack a certain sort of insight, they must also lack all awareness and free will. (One wonders where this leaves five-year-old children.)"* Mc Dermott (1995) s. 2.

[86] *„Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth"* Penrose (1994) s. 76.

bezrozporné a že historie ukazuje, že i v matematice je možné považovat nějaké tvrzení za správné, dokud se nedokáže jeho omylnost. Jiným příkladem, než uvádí McDermott, může být samotný Hilbertův program, u kterého se předpokládalo, že bude splněn až do chvíle, kdy jeho nesplnitelnost dokázal Gödel.

Matematici tudíž také nikdy neskončí svůj ověřovací proces a nerozhodnou o nějakém problému navždy[87] – mělo by se to brát jako nemožnost dokázat rozhodnutelnost tohoto problému? V tom případě by to bylo stejné, jako problém zastavení u algoritmů – pokud ano, problém by také nebyl rozhodnutelný a to ani „nevypočitatelnými" prostředky.

Současná pozice kompucionalismu (zastávána i McDermottem) naznačuje, že je třeba připustit možnosti sebe-reflexe a sebe-reference složitých inteligentních systémů, které pracují s vlastním modelem či teorií – McDermott užívá termínu „simulacrum"[88] Počítač je pak schopen začlenit sám sebe do tohoto prostředí. Do dalších detailů současného kompucionalismu se zde nebudu pouště, uzavřu McDermottovu kritiku tím, že přes četné obtíže, které výpočetní model vědomí zajisté má, je podle McDermotta schopen smysluplně zkoumat individuální poznávací schopnosti, přičemž na základě těchto poznatků někdy v budoucnu, snad, objasní i teorii inteligence obecně.[89]

---

[87] *„Human mathematicians do not generate an answer to a problem and then stop thinking about it. In fact, human mathematicians never stop."* McDermott (1995) s. 4.

[88] *„The basic idea is that a computational system can often be said to have a model or theory of some part of its environment. I hesitate to use either the word "model" or "theory" here, because of the danger that some will assume I mean to use these words in the senses they have in mathematical logic, and I emphatically do not. Perhaps "simulacrum" is the right word; some computational systems maintain simulacra of some part of their surroundings. A simulacrum allows the system to explain and predict the behavior of the world around it."* McDermott (1995) s.12.

[89] *„Computationalism is scarcely examined, let alone refuted, by this book, which stakes all its marbles on the Gödelian-gap argument, and loses. A computational theory of consciousness has many problems, but is better worked out than any alternative, including especially Penrose's. It is not arrogance, but a humble desire for truth, that leads some researchers to pursue the computational theory as a working hypothesis. The biggest obstacle to the success of this theory is not the absence of an account of conscious*

S pojmem inteligence se pojí i námitky Roberta Shanka či Marvina Minského. Roger Shank podotýká, že inteligence počítače by neměla být posuzována podle množství a obtížnosti výpočtů, které je schopen program provést a důkazů, které je schopen vysvětlit, protože myšlení z těchto typů výpočtů nemůže povstat – pokud se podíváme na lidské myšlení, uvědomíme si, že inteligence není spojena s těmito důkazy. *„Důkaz se týká toho, že aby stroj byl schopen myšlení, je nutné vyřešit ohromné množství problémů. Chyba spočívá v předpokladu, že myšlení povstává právě z těchto druhů výpočtů, o nichž je řeč. To nejspíš správné předpoklady nejsou. Vlastně všechno, co jsme se o lidském myšlení naučili, říká, že jsou zcela mylné.“*[90]

Marvin Minsky nesouhlasí s Penrosovým vysvětlením, že lidé jsou oproti počítačům schopni intuitivně řešit problémy typu zastavení Turingova stroje, či Gödelových důkazů neúplnosti. Tyto problémy jsou neřešitelné jen v tom smyslu, že neexistuje program, který by tyto problémy řešil a přitom se nikdy nemýlil. Lidská mysl však, stejně jako počítač, nemusí být chápána pouze jako dokonale a bezchybně logická.

---

awareness per se, but the fact that AI has as yet made little progress on the problem of general intelligence, and has decided to focus on a more modest strategy of studying individual cognitive skills." McDermott (1995) s. 15.

[90] Brockman (2008) s. 258.

3.3. CELKOVÉ ZHODNOCENÍ UŽITÍ GÖDELOVÝCH VĚT

Obsah diskuzí kolem možností a limitů AI, se od publikování textu J. R. Lucase, kdy se jednalo o principiální soupeření mechanicismu vs. mentalismu, velmi proměnil. Přesto, že se Penrosova argumentace snaží dokázat nevypočitatelnost lidského vědomí, v zásadě dokazuje to samé, co Lucas – nemožnost lidské vědomí simulovat pomocí počítačů. Lucas se snaží o vyvrácení mechanicismu, Penrose zase komputationismu, což pokud jde o celé spektrum možných přístupů k výpočtu a AI, je si velice blízko.

Zatímco Lucas své vyvrácení buduje vzhledem k *a-machine*, Penrose vztahuje své závěry i k *o-machine*, tedy oracle machine, což je stroj lišící se od běžného Turingova stroje i možnou přítomností nevypočitatelných funkcí, které však nemůžeme popsat. Jedním z textů, který problematizuje pojetí stroje v Lucasově a především v Penrosově argumentaci je *The Mathematical Objection: Turing, Gödel, and Penrose on the Mind* Jacka Copelanda, který upozorňuje například na následující pasáž Penrosovy knihy *SM*: „*The arguments of Part I of this book can be applied equally well against an oraclemachine model of mathematical understanding as they were against the Turingmachine model, almost without change*." [91] *O-machine*, jako zidealizovaný model, je však schopen rozhodnout problém zastavení, což je v přímém rozporu Penrosova argumentu opírajícího se o Gödelovy věty.

Od dob prvního publikování Lucasova textu a tedy i jeho námitky proti AI došlo k rozlišení mezi uzavřeným a otevřeným systémem, kde otevřený systém je schopen interakce s okolím a je schopen sebe-reflexe – Lucas pracoval s představou naprosto deterministického uzavřeného systému, jež není schopen rozhodovat jinak, než na základě základních axiomů. (Což bylo zásadní pro jeho důkaz algoritmické nedokazatelnosti Gödelových vět, čímž mohl v principu odlišit lidskou mysl od stroje.)

[91] Penrose (1994) s. 380. Uvedeno v Copeland (2008) s. 4.

Položme si proto hypotetickou otázku, pokud přijímáme Gödelovy věty jako vyvrácení možnosti inteligentního uzavřeného systému, protože není schopen dokázat svou vlastní konzistentnost, budeme schopni přijmout Gödelův důkaz provedený otevřeným systémem, dostatečně složitým a komplexním, za jasný důkaz jeho inteligence – či jak by vyplynulo z Penrosovy kritiky – přiřkneme takovému systému status nevypočitatelného vědomí?

Pokud totiž striktně užíváme určitých prostředků k vyvrácení možnosti vědomí u AI, měli bychom být připraveni přijmout důsledky, které plynou z opačného závěru, při použití stejných kritérií posuzování. *„Gödelův objev tak jistým způsobem potvrzuje, že dosáhne-li systém určitého stupně složitosti (Gödelovy teorémy platí pro formální systémy obsahující minimálně aritmetiku), pak se nutně objevují nové – emergentní – vlastnosti celku, které nelze odvodit pouze z jeho částí (tj. z axiomů a pravidel odvozování). I na tomto základě lez uvažovat o cestě ke kontextuálním strojům a přesahu od pouhé syntaxe k sémantice. Díky vnitřní síti vazeb, která zahrnuje model okolí a sebe sama jako konající entity (agenta), datového informačního toku z čidel a senzorů, které umožňují určité „chování ve světě“, schopnosti porovnávat a učit se, mohou vzniknout vnitřní reprezentace se sémantickými obsahy jako emergentní vlastnost vyšší úrovně popisu.“*[92]

Hilary Putnam ve své recenzi knihy *Shadows of the mind* uvádí, že Roger Penrose je pravděpodobně posledním zastáncem Lucasova kontroverzního způsobu vyvrácení možnosti simulovat aktivitu lidského mozku počítačem, za použití dobře známého Gödelova druhého teorému o neúplnosti. Zatímco Lucas skrze Gödelův teorém dokazoval tajuplnou podstatu lidské mysli, jež má jen málo společného s chemickými a fyzikálními vlastnostmi mozku, Penrose se snaží vykreslit zcela jiný závěr – skrze nutnost nevypočitatelných procesů v našem mozku je zapotřebí ustanovit novou vědu, která je bude moci adekvátně popsat. Penrose nejen, že užívá Gödelova teorému na obranu své vlastní verze Lucasova argumentu, ale navíc podrobně rozebírá současný stav fyziky, biologie a neurologie.

[92] Havlík (2007) s. 176.

Penrose užívá ke svému důkazu nevýpočetního charakteru vědomí Gödelův teorém, avšak způsob jeho využití byl mnohými odborníky kritizován a označen za chybný již v případě Lucase. Gödelův teorém o neúplnosti [93] ve svém zjednodušeném a zpopularizovaném znění „Pokud je S konzistentní, pak tento fakt nemůže být dokázán v S“ může být pro diskuzi o možnostech AI zavádějící. V Penrosově případě není znění teorému zjednodušeno, ale jak bylo uvedeno výše, je mnohdy zavádějící způsobem, jakým je zapojen do další interpretace problému. Je třeba si proto položit zásadní otázku, má užití Gödelova teorému relevantní hodnotu v diskuzi o možnosti či nemožnosti AI?

Domnívám se, že Gödelův teorém o neúplnosti není pro oblast AI žádným překvapením – Turing na problému rozhodnutelnosti založil svou tezi, následovanou závěrem, že neexistuje algoritmická rozhodnutelnost určitých problémů. Matematické výtky, založené na větě o neúplnosti, si by dobře vědom „*Recently the theorem of Godel and related results (Godel 1, Church 1, Turing 1) have shown that if one tries to use machines for such purposes as determing the truth or falsity of mathematical theorems and one is not willing to tolerate an occasional wrong reset, then any given machine will in some cases unable to give an answer at all. On the other hand the human intelligence seems to be able to find methods of ever increasing power for dealing with such problems „transcending“ the methods available to machines.“* [94] a také na ni náležitým způsobem odpovídá ve svém textu *Intelligent Machinery,* když říká, že neomylnost není nutnou podmínkou inteligence.[95]

---

[93] *„Gödel's Second Incompleteness Theorem states that if a system S of formalized mathematics—that is, a set of axioms and rules so precisely described that a computer could be programmed to check proofs in the system for correctness — is strong enough for us to do number theory in it, then a certain well-formed statement of the system, one which implies that the system is consistent, cannot be proved within the system.“* Putnam (1994) s. 370.

[94] Turing (1940) s. 2.

[95] *„The argument from Godel´s and other theorems rests Essentials on the condition that the machine must not make mistakes. But this is not a reguirement for intelligence.“* Turing (1940) s. 3.

Přesto nebyla Turingova odpověď na matematickou námitku obecně přijata, i když zastánci AI je respektována, co bylo důvodem jejího odmítnutí? Nepřesvědčivost, nebo obava z důsledků, které by přineslo jejího přijetí? Diskuze kolem AI stále pokračuje, což je pochopitelné vzhledem k širokému spektru možných přístupů k ní. Využití Gödelových vět pro účely této diskuze je odmítáno již téměř 50 let, a přesto nenacházíme jednoznačnou odpověď na naši otázku po korektnosti užití těchto teorémů a jejich relevantnosti v naznačené diskuzi.

Z toho důvodu se musím ptát, zda na relevantnosti Gödelových vět (ve věci vyvrácení AI) netrváme příliš zatvrzele a zda svým přístupem jejich význam nezdiskreditujeme. To, co nám říkají Gödelovy věty o neúplnosti formálních systémů, nemusí být nutně v přímém rozporu s Turingovou odpovědí na výše zmíněnou matematickou námitku.

ZÁVĚR

Práce si kladla za cíl zpřehlednit problematiku významu Gödelových vět v diskuzi o možnostech a omezeních umělé inteligence. V prvé řadě bylo proto v práci rozvedeno znění Gödelovy věty o úplnosti a znění obou vět o neúplnosti. Důraz byl ovšem kladen i na vztah mezi Gödelovými větami a Hilbertovým programem, který by z historického pohledu neměl být chápán pouze jako program, který byl skrze Gödelovy důkazy překonán, ale jako program, který umožnil definovat matematiku tak, jak ji známe dnes. Hilbertův program se snažil o úplnou axiomatizaci matematiky, což jak se ukázalo později, není možné, ale v žádném případě nebyla tato snaha nesmyslná. Ve své době byl naopak úspěch Hilbertova programu předpokládán a bez základů položených Hilbertem by Gödel své důkazy o neúplnosti pravděpodobně ani nemohl definovat.

V další části textu byla rozebrána práce Alana Turinga a její vazba nejen na Gödelovy důkazy o neúplnosti, ale také na problémy Hilbertova programu, které se Turing pokusil vyřešit novou metodou – algoritmizací daného problému. Původní snaha o nalezení výpočetního řešení problému rozhodnutelnosti, který byl Turingem shledán jako algoritmicky neřešitelný problém, však umožnila vzniknout daleko komplexnějšímu oboru – budoucí informatice a umělé inteligenci.

Ve třetí části práce je podrobně rozebrána snaha J. R. Lucase a Rogera Penrose, skrze Gödelovy teorémy, vyvrátit mechanicismus a výpočetní teorii vědomí. V práci je dále naznačena diskuze, kterou texty zmíněných autorů vyvolaly. Důraz byl kladen na příspěvky, které se zabývají užitím Gödelových teorémů v této diskuzi, např. texty Davida Lewise, Davida Boyera, Davida Codera, Solomona Fefermana či Drewa McDermotta. Hlavním záměrem práce bylo zjistit, zda jsou Gödelovy teorémy v pracích J. R. Lucase a Rogera Penrose užívány korektním způsobem, či zda je jejich znění jakkoli zatíženo dezinterpretací. K tomu se také váže ústřední otázka celé práce, a to sice, zda je užití Gödelových teorémů pro diskuzi o možnostech AI relevantní či nikoli – dokazují skutečně Gödelovy teorémy nemožnost AI tak, jak se snaží ukázat texty Lucase a Penrose?

S využitím podrobného rozboru primárních i sekundárních textů bylo v práci ukázáno, že užití Gödelových teorémů pro diskuzi kolem AI bývá odmítáno. Kritizován nebývá význam těchto vět jako takový, ale jejich užití v textech Lucase i Penrose. Mnohými kritiky bývá užití těchto vět označeno za nekorektní a také bývá tematizována míra dezinterpretace, ke které dochází v důsledku zjednodušení Gödelových vět výše jmenovanými autory.

Hlavním záměrem práce bylo zpřehlednit nastíněnou problematiku a v co největší možné míře naznačit, zda existuje řešení tohoto problému, tedy rozhodnout, zda jsou Gödelovy věty v diskuzi relevantní. Z toho důvodu bylo třeba vyřešit, k čemu konkrétně se Gödelovy věty vyjadřují a k čemu jsou naopak užívány. V neposlední řadě byla proto připomenuta matematická námitka, zakládající se na Gödelově teorému, jejíž relevanci a dopad řešil již Turing.

Domnívám se, že záměry práce byly v tomto textu splněny. Bohužel otázka po relevantnosti užití Gödelových vět v diskuzi o možnostech AI zůstává obecně nezodpovězena, protože může dojít ke zpochybnění názorů výše uvedených kritiků, kteří relevanci užití vět vyvracejí. Domnívám se však, že lze pokládat za přínosný minimálně způsob, jakým byla problematika zpracována. Problematika byla objasněna od samého základu, včetně výkladu logických východisek. Díky tomu bylo možné analyzovat texty výše jmenovaných autorů, a jejich kritiků, a v neposlední řadě zvážit relevanci těchto textů. Za původní vklad, kromě metody zpracování problematiky, můžeme považovat i volbu analyzovaných textů a především zapojení jejich hlavních tezí do vlastní argumentace.

ANOTACE

Filosofická fakulta Univerzity Palackého v Olomouci

Katedry filosofie

Autor práce: Martina Juříková

Vedoucí práce: prof. Jan Štěpán

Název diplomové práce: KURT GÖDEL A PROBLEMATIKA UMĚLÉ INTELIGENCE

Počet znaků: 118 496

Počet příloh: 0

Počet titulů použité literatury: 37

Klíčová slova: Kurt Gödel, Věta o neúplnosti, Alan Turing, Rozhodnutelnost, J. R. Lucas, Roger Penrose, Umělá inteligence.

Anotace:

Diplomová práce „Kurt Gödel a problematika umělé inteligence“ se zabývá rozborem Gödelových vět o neúplnosti, jejich vazbou a vlivem na práci Alana Turinga a jejich významem pro oblast umělé inteligence. Konkrétně pak řeší korektnost interpretace těchto vět v argumentaci J. R. Lucase a Rogera Penrose. V neposlední řadě se skrze analýzu dalších textů snaží zodpovědět otázku, zda jsou Gödelovy věty relevantním argumentem v diskuzi o umělé inteligenci.

Název diplomové práce v anglickém jazyce: KURT GÖDEL AND THE PROBLEMS OF ARTIFICIAL INTELLIGENCE

Klíčová slova v anglickém jazyce: Kurt Gödel, Incompleteness theorem, Alan Turing, J. R. Lucas, Roger Penrose, Artificial Intelligence

Anotace v anglickém jazyce:

This thesis called „Kurt Gödel and the problems of the Artificial Intelligence“ deals with the analysis of Gödel´s theorems of incompleteness, their connection and influence on the work of Alan Turing as well as their importance to the field of the Artificial Intelligence. To be more specific the work is solving correctness of the interpretation of these theorems in argumentation of J. R. Lucas and Roger Penrose. At last but not least case, using the analysis of other texts we are trying to answer the question whether or not Gödel´s theorems are relevant to discussion about the Artificial Intelligence.

POUŽITÁ LITERATURA:

Barrow, John D. (2000): *Pí na nebesích: O počítání, myšlení a bytí*. Mladá fronta, Praha.

Bečvář, Jiří (1971): *O druhém Hilbertově problému*. In: Pokroky matematiky, fyziky a astronomie, vol. 16/ No. 5. 225 – 237.

Běhounek, Libor (2006): *Kurt Gödel: život, výsledky a jejich význam*. In: Kognice a umělý život VI. Slezská univerzita v Opavě, Opava, s. 47 – 57.

Berka, Petr (2008): *Inteligentní systémy*. Oeconomika, Praha.

Boyer, David L. (1983): *J.R. Lucas, Kurt Gödel, and Fred Astaire*. In: The Philosophical Quarterly, vol. 33/ No. 131. 147 – 159. http://www.jstor.org/stable/2218741

Brockman.John (2008): *Třetí kultura: za hranice vědecké revoluce*. Academia, Praha.

Coder, David (2003): *Gödel's Theorem and Mechanism*. In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Copeland, Jack (2008): *The Mathematical Objection: Turing, Gödel, and Penrose on the Mind*. Conference paper: Computation and Cognitive Science, King's College Cambridge, 2008. http://people.pwf.cam.ac.uk/mds26/cogsci/files/Copeland---TheMathematicalObjection.pdf

Coveney, Peter V. (2003): *Mezi chaosem a řádem: hranice komplexity: hledání řádu v chaotickém světě*. Mladá fronta, Praha.

Feferman, Solomon (1995): *Penroe´s Gödelian argument*. In: Psyche, vol. 2/ No. 7. http://math.stanford.edu/~feferman/papers/penrose.pdf

Goldsteinová, Rebecca (2005): *Neúplnost: Důkaz a paradox Kurta Gödela*. Argo, Praha.

Good, I. J. (2003): *Human and Machine Logic*. In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Gödel, Kurt (1931): *O formálně nerozhodnutelných větách v díle Principia Mathematica a příbuzných systémech*. In: Novotný, Malina (ed.): Kurt Gödel. NAUMA, Brno. 1996.

Havlík, Vladimír (2007): *Kurt Gödel a AI*. In: Meze formalizace, analytičnosti a prostoročasu. Filosofia, Praha, 161 – 177.

Jančar, Petr (2007): *Úvod do teoretické informatiky*. VŠB-TUO. Ostrava.

Kolman, Vojtěch (2008): *Filosofie čísla*. Filosofia, Praha.

Leavitt, David (2007): *Muž, který věděl příliš mnoho*: Alan Turing a první počítač. Argo, Praha.

Lewis, David (2003): *Lucas against Mechanism*. In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Lucas, John R. (2003): *Human and Machine Logic: a Rejoinder*. In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Lucas, John R. (2003): *Lucas against Mechanism: a Rejoinder.* In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Lucas, John R. (2003): *Minds, Machines and Gödel*. In: Etica & Politica, vol. 5/ No. 1. http://www2.units.it/etica/2003_1/index.html

Lukasová, Alena (1995): *Logické základy umělé inteligence*: Výroková a predikátová logika 1. Ostravská univerzita, Ostrava.

Lukasová, Alena (2003): *Formální logika v umělé inteligenci*. Computer Press, Brno.

Malina, Jaroslav, Novotný, Jan, ed.(1996): *Kurt Gödel*. NAUMA, Brno.

McDermott, Drew (1995): *Penrose is Wrong*. In: Psyche, vol. 2/ No. 17. (http://www.calculemus.org/MathUniversalis/NS/10/09mcdermott.html)

Nagel, Ernest, Newman, James R. (2006): *Gödelův důkaz*. VUTIUM, Brno.

Penrose, Roger (1994): *Shadows of the Mind*. Oxford University Press, New York.

Penrose, Roger (1999): *Makrosvět, mikrosvět a lidská mysl*. Mladá fronta, Praha.

Peregrin, Jaroslav (2005): *Kapitoly z analytické filosofie*. Filosofia, Praha.

Peterka, Jiří (1994): *Problém zastavení*. In: Computerworld No. 33. http://www.earchiv.cz/a94/a433c120.php3

Putnam, Hilary (1995): *Book Reviews: Shadows of the Mind*. In: Bulletin of the American Mathematical Society, vol. 32/ No. 3. 370 – 373. (http://www.ams.org/journals/bull/1995-32-03/S0273-0979-1995-00606-3/S0273-0979-1995-00606-3.pdf)

Searle, John R. (1980): *Minds, Brains and Programs*. In: The Behavioral and Brain Sciences, vol. 3/ No. 3, 417 – 424. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.5248&rep=rep1&type=pdf

Smullayn, Raymond (2003): *Navěky nerozhodnuto*. Academia, Praha.

Turing, Alan (1937): *On computable numbers, with an application to the Entcheidungsproblem.* Dostupné z Turingarchive.org / AMT/B/12 http://www.turingarchive.org/browse.php/B/12

Turing, Alan (1940): *Intelligent Machinery.* Dostupné z Turingarchive.org / AMT/C/11 (http://www.turingarchive.org/browse.php/C/11)

Turing, Alan (1950): *Computing Machinery and Intelligence*. In: Mind, vol. 59/ No. 236, 433 – 460. http://blog.santafe.edu/wp-content/uploads/2009/05/turing1950.pdf

Zlatoš, Pavol (2007): *Ani matematika si nemože by istá sama sebou*. IRIS, Bratislava.