



Uniwersytet Łódzki
Wydział Matematyki i Informatyki
Kierunek Informatyka

Urszula Jędrzejczak
289529

**„CZY MASZYNY MOGĄ MYŚLEĆ?”:
ROZWAŻANIA NA TEMAT SILNEJ
SZTUCZNEJ INTELIGENCJI**

PRACA MAGISTERSKA

Wykonana pod kierunkiem
doktora Piotra Fulmańskiego

Łódź, 2013

SPIS TREŚCI:

WSTĘP	2
I. CZY OSOBLIWOŚĆ JEST BLISKO?	3
I.1. PRZEWIDYWANIA	3
I.2. AI DZIŚ	6
<i>I.2.1. Symulacje mózgu</i>	6
<i>I.2.2. Wyszukiwanie, rozumowanie, bazy wiedzy</i>	7
<i>I.2.3. Rozpoznawanie mowy</i>	12
<i>I.2.4. Tłumaczenie maszynowe</i>	12
<i>I.2.5. Twórczość</i>	13
<i>I.2.6. Robotyka</i>	14
II. CZY MASZYNY MOGĄ MYŚLEĆ?	16
II.1. PRAWO MOORE'A I PRAWO KURZWEIL'A	18
II.2. ZARZUTY, SPRZECIWY, ARGUMENTY... ..	21
<i>II.2.1. Computing Machinery and Intelligence</i>	21
<i>II.2.2. Inne argumenty</i>	31
III. CZY MASZYNY POWINNY MYŚLEĆ?	38
III.1. CZŁOWIEK 2.0?	40
<i>III.1.1. Czemu przyszłość nas nie potrzebuje</i>	40
<i>III.1.2. Passing the Torch</i>	43
III.2. PROBLEMY PRAKTYCZNE	45
<i>III.2.1. Trzy prawa robotyki</i>	45
<i>III.2.2. Kto osądzi sztuczną inteligencję?</i>	49
<i>III.2.3. Race against the machine</i>	52
IV. THE IMITATION GAME	56
IV.1. TEST NA CZŁOWIECZEŃSTWO	57
<i>IV.1.1. Definicja</i>	57
<i>IV.1.2. Wersja współczesna</i>	59
IV.2. SZTUKA PODSTĘPU	64
<i>IV.2.1. ELIZA</i>	64
<i>IV.2.2. Chatboty</i>	69
PODSUMOWANIE:	81
BIBLIOGRAFIA:	82
SPIS ILUSTRACJI:	86

WSTĘP

Sztuczna inteligencja przeżywa obecnie renesans swojego rozwoju, w dużej mierze za sprawą stosowania metod opartych na statystyce oraz przetwarzaniu wielkich ilości danych („Big Data”). Podobnie jak na początku historii tej dziedziny informatyki, sukcesy stają się źródłem zarówno optymistycznych, jak i katastroficznych prognoz dotyczących przyszłości technologii. Szczególnie kontrowersyjnym i ciągle obecnym w zbiorowej wyobraźni tematem pozostaje kwestia możliwości powstania sztucznej inteligencji ogólnego zastosowania.

W 1950 roku Alan Turing, w słynnym tekście *Computing Machinery and Intelligence* podjął się próby odpowiedzi na pytanie „Czy maszyny mogą myśleć”, na płaszczyźnie logicznej, filozoficznej i technicznej. Wymyślił również test mający sprawdzać inteligencję maszyn, nazwany później testem Turinga, oraz zaproponował prognozy dotyczące czasu, w którym jego zdaniem maszynom uda się spełnić postawione przez niego warunki. Celem niniejszej pracy jest ponowne rozpatrzenie klasycznego pytania „Czy maszyny mogą myśleć” na wymienionych wyżej płaszczyznach oraz prześledzenie rozwoju technologii i myśli, jakie towarzyszyły próbom stworzenia myślących maszyn.

Pierwszy rozdział zawiera omówienie tego jak wiele zmieniło się w kwestii przewidywań dotyczących stworzenia myślących maszyn w ciągu ponad sześćdziesięciu lat istnienia sztucznej inteligencji oraz jak wiele udało się w tym czasie osiągnąć. Rozdział drugi poświęcony jest teoretycznym możliwościom powstania sztucznej inteligencji ogólnego zastosowania. Rozdział trzeci poszerza tematykę rozdziału drugiego o dyskusję ewentualnych konsekwencji jej powstania, natomiast rozdział czwarty zawiera omówienie testu Turinga oraz jego współczesnej wersji.

I. CZY OSOBLIWOŚĆ JEST BLISKO?¹

I.1. PRZEWIDYWANIA

"In 30 years, it would be as easy to ask a computer a question as to ask a person."

Alan Turing

"Za 30 lat zadanie pytania komputerowi, będzie tak proste, jak zadanie pytania człowiekowi"² - to założenie, sformułowane przez Alana Turinga, nadal nie znajduje potwierdzenia w rzeczywistości, mimo, że od wypowiedzenia tych słów mija w tym roku 76 lat. Marvin Minsky (współtwórca laboratorium AI na MIT) w 1967 roku przewidywał, że „W przeciągu dwóch dekad... niewiele dziedzin intelektu pozostanie poza rzeczywistością maszyn – problem stworzenia sztucznej inteligencji będzie zasadniczo rozwiązany”³. Z kolei Herbert Simon w 1965 roku stwierdził, że “w ciągu 20 lat maszyny będą zdolne do każdej pracy, którą może wykonać człowiek”⁴.

Dziś wiadomo już, że przewidywania ojców sztucznej inteligencji, futurologów, czy wreszcie autorów science fiction z ubiegłego wieku okazały się być przesadnie optymistyczne. Pomimo pojawiających się od czasu do czasu prasowych rewelacji, współczesnej nauce nie udało się dotychczas stworzyć programu, czy też maszyny, której udałoby się dorównać człowiekowi choćby w sztuce prowadzenia

¹ Tytuł jest nawiązaniem do tytułu książki Raymonda Kurzweila: *Singularity Is Near: When Humans Transcend Biology*

² [tłumaczenie własne]

³ “Within a generation... few compartments of intellect will remain outside the machine’s realm – the problem of creating »artificial intelligence« will be substantially solved”

Marvin Lee Minsky, *Computation: Finite and Infinite Machines*, New Jersey: Prentice-Hall, 1967 [tłumaczenie własne]

Trzeba zaznaczyć, że obecne przewidywania M.Minsky’ego są nieco mniej optymistyczne

⁴ “machines will be capable, within twenty years, of doing any work that a man can do” [tłumaczenie własne]

Herbert Simon, *The Shape of Automation for Man and Management*, New York: Harper&Row, 1965

konwersacji (możliwość swobodnej rozmowy z komputerem zakorzeniła się w zbiorowej świadomości jako uniwersalny test inteligencji maszyny za sprawą testu zaproponowanego ponad 50 lat temu przez Alana Turinga; historia i krytyka testu Turinga zostanie bardziej szczegółowo omówiona w następnym rozdziale). Nie oznacza to oczywiście, że w dziedzinie sztucznej inteligencji nie nastąpił żaden znaczący postęp. Należy jednak wprowadzić rozróżnienie na tzw. „słabą” oraz „silną” sztuczną inteligencję.

Ogólnie można powiedzieć, że wszelkie istniejące w tej chwili programy AI są przedstawicielami „słabej” sztucznej inteligencji, a więc takiej, która jedynie zachowuje się jak gdyby była inteligentna, dodatkowo tylko w wąskim zakresie (np. system ekspertowy nie zagra z nami w szachy). Natomiast „silna” sztuczna inteligencja to taka, o której możemy powiedzieć, że działa inteligentnie, ponieważ sama myśli i jest zdolna konkurować z inteligencją człowieka. Termin ten zaproponował po raz pierwszy filozof John Searle w 1980 roku.⁵

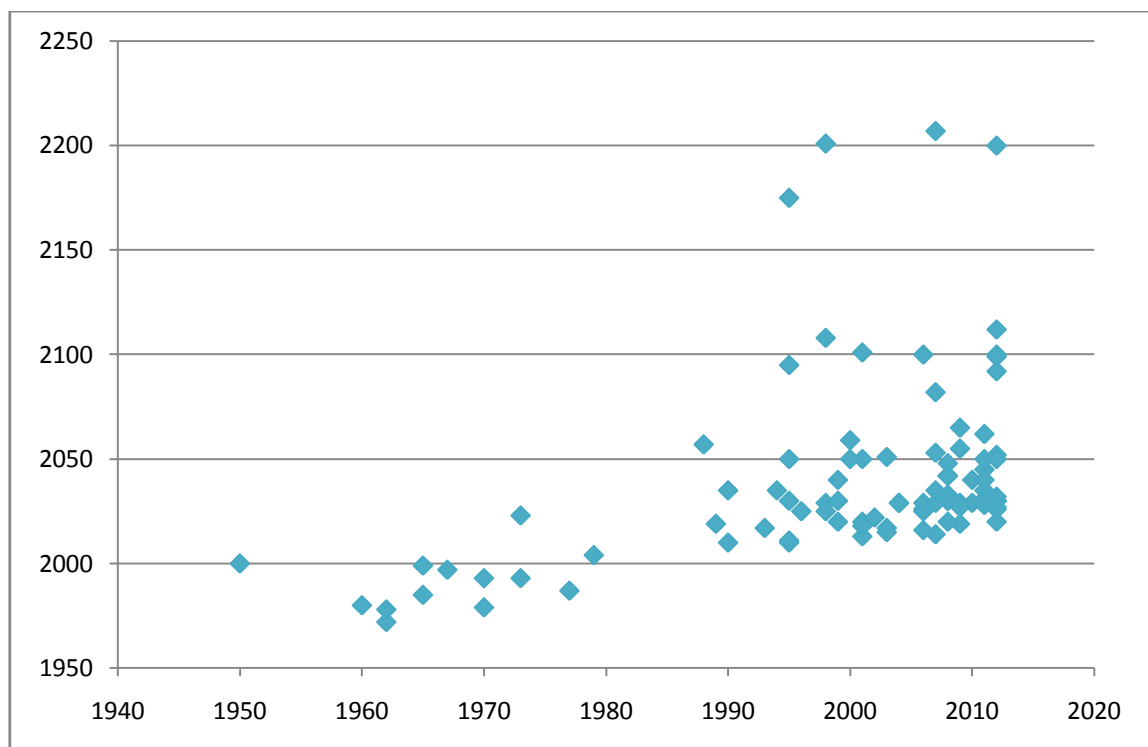
Stawianie stworzenia myślącej i działającej jak człowiek „silnie” sztucznej inteligencji (SSI) jako cel tej gałęzi nauki było wielokrotnie krytykowane i przeciwstawiane innym podejściom (w których celem sztucznej inteligencji jest stworzenie maszyn rozumujących lub działających racjonalnie, a nie imitujących ludzkie zachowania). Nierozstrzygnięta pozostaje też kwestia, czy stworzenie SSI jest w ogóle możliwe. Przykładowo optymistyczne poglądy m.in. M.Minsky’ego były ostro krytykowane już w 1972 roku przez Huberta Dreyfusa w książce *What Computer Still Can't Do: A Critique of Artificial Reason*.

Z drugiej strony, niektórzy nie poprzestają na SSI. Ray Kurzweil (f futurolog, transhumanista, obecny Director of Engineering w Google), przewiduje, że powstanie „silnej” sztucznej inteligencji zapoczątkuje tzw. „osobliwość technologiczną” (*singularity*), czyli osiągnięcie przez systemy komputerowe inteligencji przekraczającej inteligencję całej ludzkości. Osobliwość to moment, w którym powstanie inteligentnych maszyn zdolnych do samoprogramowania się sprawi, że postęp technologiczny wymknie nam się z rąk i nowa technologia będzie poza zasięgiem ludzkiego zrozumienia. W związku z tym wszelkie przewidywania

⁵ Definicje silnej sztucznej inteligencji nieco się od siebie różnią, w dalszej części tekstu będę używać tego pojęcia w rozumieniu sztucznej inteligencji ogólnego zastosowania

dotyczące świata „po osobliwości” mogą być jedynie spekulacją. Niektóre szacunki zakładają możliwe wystąpienie osobliwości przed rokiem 2100⁶. Co ciekawe, teoria osobliwości niekoniecznie zakłada, że powstające superinteligentne maszyny będą posiadały świadomość.

Zróznicowanie przewidywań co do ewentualnego powstania silnie sztucznej inteligencji podsumowuje wykres przedstawiający rok przewidywanego wystąpienia SSI w zależności od roku przewidywania (dane obejmują przewidywania ekspertów, jak Simon, Minsky i Turing, oraz dziennikarzy, pisarzy i futurologów, jak V.Vinge). Dane pochodzą z materiałów The Singularity Institute⁷



⁶ Luke Muehlhauser, Anna Salamon, *Intelligence Explosion: Evidence and Import* [online]. [dostęp: 7.02.2013]. Dostępny: <http://intelligence.org/files/IE-EI.pdf>

⁷ Stuart Armstrong, Kaj Sotala: *How We're Predicting AI—or Failing To* [online]. [dostęp: 7.02.2013]. Dostępny: <http://intelligence.org/files/PredictingAI.pdf>

I.2. AI DZIŚ

„Co to znaczy – artificial servility? Takiego słowotworu nie było, dopóki go niniejszym nie urobiłem. Jest to, w literalnym tłumaczeniu »sztuczne niewolnictwo«, a dotyczy wszystkiego, co potrafią obecnie tak rozpowszechniające się na całym świecie elektroniczne urządzenia do przetwarzania, uruchamiania, przechowywania i last but not least do przekazywania informacji. Dlaczego »niewolnictwo«? Dlatego, ponieważ w całym owym przemyśle (przynoszącym różnym Microsoftom miliardy), w całym tłumie komputerowym, we wszystkich generacjach hardware i software, modemach, serwerach nie ma śladu rozumu. Nie ma ani krzty inteligencji. Pracują tak, jak niewolnicy: na nasz rozkaz.”

Stanisław Lem, Artificial Servility

Mimo, że wczesne nadzieje związane ze stworzeniem sztucznej inteligencji dorównującej człowiekowi okazały się przedwczesne oraz pomimo głosów krytyki, stworzenie SSI pozostaje marzeniem wielu badaczy AI i futurologów.

I.2.1. SYMULACJE MÓZGU

Obecnie, jednym z największych i najgłośniejszych projektów dotyczących AI jest Human Brain Project, którego celem jest zasymulowanie działania całego mózgu, neuron po neuronie. Badacze mają nadzieję poznać w ten sposób działanie mózgu metodą bottom-up Tym, co odróżnia HBP od poprzednich tego typu badań, jest to, że grupa badaczy pod kierownictwem Henry’ego Makrama chce dokładnie zasymulować komórki neuronów (z uwzględnieniem ich wielkiej różnorodności) z drobiazgową dokładnością, bez żadnych uproszczeń (włącznie ze strukturą kanałów jonowych etc.). HBP jest następcą poprzedniego projektu Makrama, Blue Brain, w ramach którego udało się przy pomocy następcy superkomputera IBM

Blue Gene, zasymulować 100 kolumn neuronalnych mózgu szczura. Szacuje się, że do zasymulowania mózgu człowieka będzie potrzebna moc obliczeniowa sięgająca exaflopa (10^{18}), w czym z pewnością pomoże fakt, że 23 stycznia 2013 roku HBP otrzymał od Unii Europejskiej grant w wysokości 1 miliarda euro.⁸ Według przewidywań Makrama, do zbudowania pełnej symulacji ludzkiego mózgu potrzeba ok. 10 lat.

Zupełnie inne podejście do projektowania mózgu reprezentują naukowcy z Uniwersytetu Stanforda pod przewodnictwem K. Boahema, zaangażowani w projekt Neurogrid. Uważają oni, że poleganie na nieprawdopodobnie energochłonnych superkomputerach to ślepa uliczka (mózg zużywa jedynie ok. 10 Wattów), natomiast to, co jest potrzebne, to tranzystorowe odwzorowanie mózgu (przepływ jonów ma być odwzorowany bezpośrednio poprzez przepływ elektronów), o charakterze sieci. Oznacza to odejście od architektury neumannowskiej. Do poprzednich osiągnięć zespołu należy zasymulowanie w analogiczny sposób neuronów siatkówki oka.

Do innych ostatnich symulacji mózgu należą Spaun (symulacja obejmująca 2,5 miliona neuronów, zdolna do rozwiązywania prostych problemów arytmetycznych i logicznych) i SyNAPSE (symulacja IBM obejmująca 530 milionów neuronów i 100 trylionów synaps).

I.2.2. WYSZUKIWANIE, ROZUMOWANIE, BAZY WIEDZY

Przedsięwzięciem zupełnie innego rodzaju, ale również wyznaczającym granice tego, do czego jest obecnie zdolna AI, jest superkomputer Watson (nazwany tak na cześć jednego z założycieli IBM, Thomasa Watsona). Od czasu słynnej wygranej Deep Blue z Garrym Kasparowem, IBM szukał kolejnego prestiżowego wyzwania. Wybór padł na znany amerykański teleturniej telewizyjny *Jeopardy!* (w Polsce znany jako *Va Banque*), o oglądalności sięgającej średnio 25 milionów widzów. Zadaniem uczestników programu jest układanie pytań do odpowiedzi podawanych przez prowadzącego, przykładowo:

⁸ M. Mitchell Waldrop, *Computer modelling: Brain in a box*, Nature, 2012, t. 482 nr 7386, s. 456–458

Wskazówka: -When this diplomat met the singer Psy in 2012, he said, "Until 2 days ago... I was the most famous Korean in the world"

Odpowiedź: - Who is Ban Ki-moon?⁹

(Wskazówka: Kiedy ten dyplomata spotkał piosenkarza Psy w 2012 roku, powiedział: „Przed dwoma dniami... byłem najbardziej znanym Koreańczykiem na świecie” Odpowiedź: Kim jest Ban Ki-moon?¹⁰)

Wskazówka: - "Classic candy bar that's a female Supreme Court justice"

Odpowiedź: - "What is Baby Ruth Ginsburg?"¹¹

(Wskazówka:- „Klasyczny batonik, który jest kobietą sędzią Sądu Najwyższego”
Odpowiedź – „Kim jest Baby Ruth Ginsburg?”¹²

Mamy tu do czynienia z grą słów - Ruth Ginsburg jest sędzią Sądu Najwyższego Stanów Zjednoczonych, a Baby Ruth to nazwa popularnego w Stanach batonika)

Dodatkowym utrudnieniem jest to, że uczestnicy odpowiadają pod presją czasu, najlepsi gracze podejmują decyzje o udzieleniu odpowiedzi w przeciągu ułamków sekund. Gra w *Jeopardy!* wymaga szybkości i szerokiej wiedzy, obejmującej również ciekawostki, anegdoty oraz ostatnie wydarzenia, rozpoznawania gier słownych i aluzji. Watson dysponuje mocą 750 serwerów IBM POWER i szesnastoma terabajtami pamięci, jednak nawet takie zasoby nie wystarczyłyby do pokonania ludzkich przeciwników, gdyby nie nowatorskie podejście do problemu wyszukiwania informacji (Watson podczas gry nie dysponuje połączeniem do Internetu, musi więc polegać na własnych zasobach - bazie danych, będącej ekwiwalentem ok. miliona książek). Badacze pod przewodnictwem Davida Ferrucci wykorzystali ponad 100 różnych algorytmów do analizy języka, identyfikowania źródeł, znajdowania i generowania hipotez, przesłanek oraz następnie odpowiedniej oceny powstałych wcześniej hipotez¹³.

⁹ Źródło: Strona internetowa teleturnieju Jeopardy.[online].[dostęp: 7.02.2013]. Dostępny: <http://www.jeopardy.com/gamesandmobile/clueoftheday/>

¹⁰ [tłumaczenie własne]

¹¹ Clive Thompson, *What Is I.B.M.'s Watson?*.[online].[dostęp: 7.02.2013]. Dostępny: <http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html?pagewanted=1&r=3&>

¹² [tłumaczenie własne]

¹³ David Ferrucci et al., *Building Watson: An Overview of the DeepQA Project*, Ai magazine, s. 68

Mechanizm działania Watsona opiera się na statystycznej analizie danych, przetwarzaniu współbieżnym oraz odpowiednim zbalansowaniu zastosowania płytkiej i głębokiej semantyki. Watson po zaopatrzeniu go w podstawowy zestaw słowników, książek, thesaurusów, encyklopedii etc. jest w stanie sam uzupełniać już posiadane informacje, samodzielnie oceniając ich przydatność, bazując przy tym na informacjach pochodzących z Internetu. Po wstępnej analizie pytania, Watson tworzy hipotezy, czyli zestaw możliwych odpowiedzi, dla których następnie szuka uzasadnienia. Ostateczna odpowiedź jest wybierana w oparciu o prawdopodobieństwo obliczone na podstawie porównania efektów działania różnych algorytmów równoległe poszukujących najlepszej odpowiedzi. W 2011 roku Watson po serii wcześniejszych sukcesów w *Jeopardy!* zwyciężył z Bradem Rutterem oraz Kenem Jenningsem, wcześniej niepokonanym mistrzem tego programu, tym samym pieczętując swoją wyższość nad ludzkimi graczami.

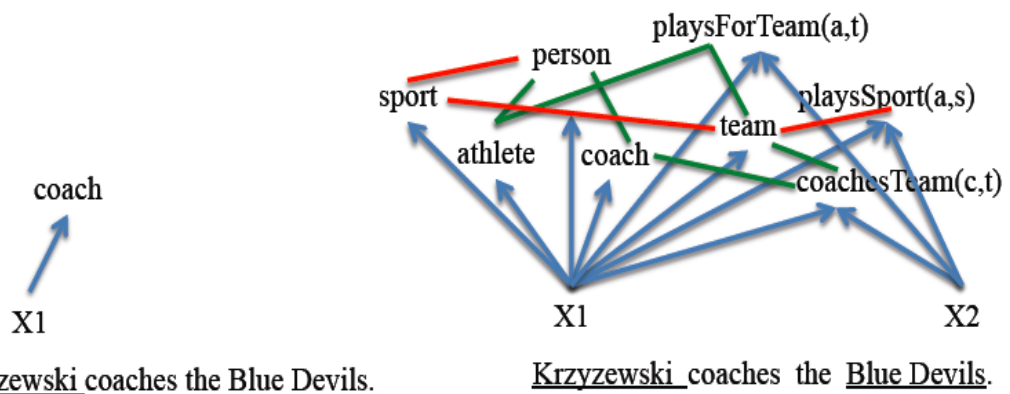
Watson (a właściwie kryjąca się za nim architektura DeepQA) nie jest jedynym rozwijanym obecnie systemem mającym za zadanie usprawnić proces wyszukiwania i wyłuskiwania informacji. Do innych przykładów należą m.in. NELL (Never-Ending Language Learner), rozwijany na Carnegie Mellon University i działający od 1984 roku projekt Cyc (Cycorp, Austin, Texas). Oba przedsięwzięcia mają na celu stworzenie kompletnej bazy wiedzy (o charakterze semantycznym, a nie syntaktycznym), na której będą mogły opierać swoje rozumowanie programy AI.

NELL jest systemem, który nieprzerwanie przeszukuje Internet, ucząc się nowych faktów, a jednocześnie cały czas samodoskonalać proces uczenia. Od roku 2010 NELL zgromadził 1,935,958 założeń i liczba ta nieprzerwanie rośnie. Baza wiedzy gromadzona przez NELL jest dostępna na stronie projektu (<http://rtw.ml.cmu.edu>).

NELL został początkowo wyposażony w ontologię definiującą ponad 600 kategorii (person, sportsTeam, fruit, emotion) i powiązań (playsOnTeam(athlete,sportsTeam), playsInstrument(musician,instrument))¹⁴ wraz z przykładami, a sam codziennie przeszukuje 100,000 stron internetowych, automatycznie poszerzając swój początkowy zestaw przekonań. NELL uczy się

¹⁴ Przykłady pochodzą ze strony projektu: <http://rtw.ml.cmu.edu/rtw/overview> , [dostęp 9.02.2012]

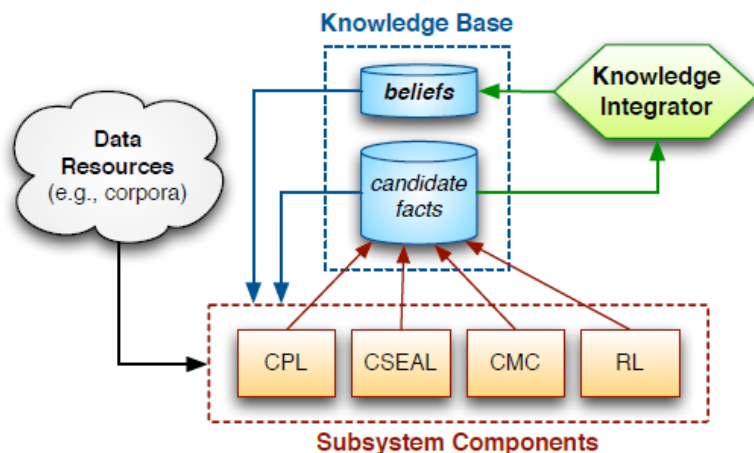
bootstrapowo, przy użyciu czterech różnych systemów wyluskiwania nowych propozycji przekonań, w założeniu popełniających nieskorelowane błędy. Dodatkowo system grupuje kategorie i relacje w taksonomie. Tym, co wyróżnia NELL'a, jest zastosowanie parowania treningu pojęć i relacji z powiązanych kategorii, co zwiększa ilość ograniczeń nakładanych na poszczególne wartości, jednocześnie zwiększając niezawodność systemu. Ideę treningu parowanego przedstawia poniższy schemat¹⁵:



Ilustracja I - 1: Idea treningu parowanego, Źródło: Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka Jr., Tom M. Mitchell: *Coupled Semi-Supervised Learning for Information Extraction*, str. 1

Nauka NELLA jest nadzorowana przez ludzi jedynie przez ok. 5 minut dziennie.

Ogólny proces przyswajania nowych informacji przez NELL'a przedstawia poniższy wykres:

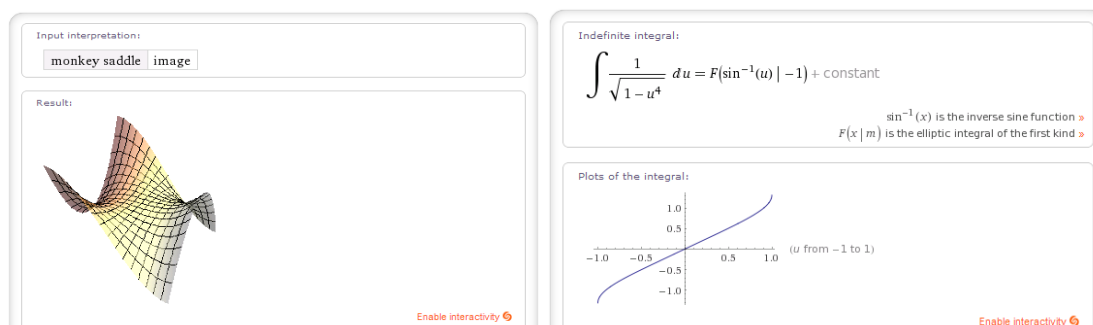


Ilustracja I - 2: Proces przyswajania informacji przez NELL'a Źródło: Andrew Carlson, Justin Betteridge, Bryan Kiesel, Burr Settles, Estevam R. Hruschka Jr., Tom M. Mitchell: *Toward an Architecture for Never-Ending Language Learning*, str 2.

Z kolei projekt rozwijany przez Cycorp opierający się na danych wprowadzonych manualnie przez ludzi, przeszukujących gazety, encyklopedie itp. w poszukiwaniu faktów. Doug Lenat zakładał, że potrzeba ok. 100 milionów założeń, zanim CYC będzie w stanie kontynuować naukę samodzielnie (wprowadzenie ich miało łącznie zająć 2 osobowiki).

Ilość produkowanych obecnie przez ludzkość informacji zaczyna przybierać rozmiary powodzi. Projekty takie jak Watson, NELL i Cyc, mają szansę pomóc w ich katalogowaniu i wyszukiwaniu. Może to być niezwykle przydatne nie tylko dla naukowców, ale również dla zwykłych ludzi, raz na zawsze zmieniając sposób wyszukiwania informacji w Internecie.

Do warty wspomnienia istniejących rozwiązań mających ułatwić wyszukiwanie poprzez interpretację zapytań formułowanych w języku naturalnym, albo chociaż zbliżonym do naturalnego, warto wspomnieć o Wolfram'ie Alpha (<http://www.wolframalpha.com/>), określanym przez jego autorów jako *computational knowledge engine*. Wolfram|Alpha potrafi udzielić odpowiedzi na pytanie „Which months have 31 days?” („które miesiące mają 31 dni?”), ale również przeprowadzać bardziej skomplikowane obliczenia, czy rysować wykresy, przykładowo poniżej przedstawiono fragmenty wyników dla zapytań „plot monkey saddle” i „integrate 1/sqrt(1-u^4)” :



Ilustracja I - 3: Przykład działania Wolfram|Alpha

Niektóre efekty działania Wolfram|Alpha są spektakularne, jednak opiera się on na bazie danych zarządzanej przez ludzi, nieustrukturyzowanej w sposób pozwalający na „rozumienie” zapytań, a także nie do końca dobrze radzi sobie z interpretacją mniej sformalizowanego języka.

1.2.3. ROZPOZNAWANIE MOWY

Gdyby naukowcom pracującym nad projektami w rodzaju Watsona i NELLA udało się osiągnąć sukces, to można by było uznać, że droga do prowadzenia rozmów z maszyną bez pomocy klawiatury jest otwarta, ponieważ systemy rozpoznawania mowy już jakiś czas temu osiągnęły poziom zaawansowania pozwalający na ich komercyjne zastosowanie. Przykładem może tu być telefoniczny system informacji lotniczej i rezerwowania lotów wprowadzony przez United Airlines już w 1999 roku.

Dodatkowo obecnie ludzie przemawiający do swoich elektronicznych gadżetów stają się powszechnym widokiem za sprawą aplikacji takich jak Siri dla Apple's iOS. Android i Windows Phone również oferują swoim użytkownikom interfejsy głosowe.

1.2.4. TŁUMACZENIE MASZYNOWE

Kolejnym wyzwaniem związanym z prowadzeniem rozmów z maszynami jest tłumaczenie automatyczne. Prawdopodobnie najbardziej znaną aplikacją tego typu jest dostępny za darmo Google Translate (<http://translate.google.com/>), który jest w stanie dokonywać tłumaczeń w 65 językach z oszałamiającą prędkością. O ile jednak GT radzi sobie doskonale ze stosunkowo prostymi zdaniami...

Zdanie wejściowe:

"Każda wystarczająco zaawansowana technologia jest nieodróżnialna od magii."

Wynik tłumaczenia na angielski:

"Any sufficiently advanced technology is indistinguishable from magic."

Wynik tłumaczenia z powrotem na język polski:

"Każda wystarczająco zaawansowana technologia jest nie do odróżnienia od magii."

...to po nieco trudniejszych przykładach widać wyraźnie, że daleko mu jeszcze do perfekcji:

Zdanie wejściowe:

"Mowa to dom bycia. W domostwie mowy mieszka człowiek. Stróżami tego domostwa są myśliciele i poeci."

Wynik tłumaczenia na angielski:

"Language is the house of being. Speech in household living man. Guardians of the homes are thinkers and poets."

Wynik tłumaczenia z powrotem na język polski:

"Język jest domem bytu. Mowa w żywego człowieka domowymi. Opiekunowie domów są myśliciele i poeci".

Do bardziej spektakularnych osiągnięć tłumaczenia maszynowego należy zaprezentowany przez Microsoft w 2012 roku program tłumaczący mówiony język angielski na chiński, zachowujący elementy stylu mówiącego. Tłumaczenie odbywa się z niewielkimi opóźnieniami w czasie rzeczywistym.

1.2.5. TWÓRCZOŚĆ

Ludzie powoli przyzwyczajają się do automatycznych tłumaczy i systemów obsługi klienta, godzą się na wyższość programów w obliczeniach matematycznych i grach logicznych. Jednak co by było gdyby AI wyszła na prowadzenie również w dziedzinie twórczości artystycznej? Wbrew pozorom nie jest to pomysł zupełnie oderwany od rzeczywistości. Dowodem na to jest sukces projektu *Experiments in Musical Intelligence* autorstwa Davida Cope'a z University of California w Santa Cruz. Dotknięty blokiem artystycznym Cope stworzył oprogramowanie zdolne analizować twórczość ludzkich kompozytorów, destylować z nich najbardziej charakterystyczne elementy i na tej podstawie tworzyć muzykę w stylu danego artysty. EMI udało się oszukać publiczność imitując kompozycje Bacha podczas eksperymentu przeprowadzonego w 1997 roku.

Ciekawym zjawiskiem jest też Music Xray (www.musicxray.com), który analizuje muzykę pod kątem właściwości akustycznych i występujących w niej wzorców matematycznych i na tej podstawie ocenia potencjał danego utworu do stania się kolejnym hitem list przebojów.

1.2.6. ROBOTYKA

Przechodząc do fizycznych aspektów sztucznej inteligencji, można powiedzieć, że roboty powoli stają się częścią naszej codzienności, m.in. za sprawą samobieżnego odkurzacza Roomba (iRobot Corporation). Trwają również prace nad robotami bardziej przypominających ich filmowe wyobrażenia, jak stworzony przez Hondę ASIMO (*Advanced Step in Innovative Mobility*). ASIMO potrafi nie tylko chodzić jak człowiek i wspinać się po schodach, ale również rozpoznawać twarze, głosy i gesty. Docelowo ASIMO ma stać się uniwersalnym domowym pomocnikiem, szczególnie dla ludzi starszych czy niepełnosprawnym.



Ilustracja 1 - 4: Robot Asimo. Źródło: materiały prasowe Hondy.
<http://asimo.honda.com/gallery/>

Wydaje się jednak, że stworzenie godnego zaufania robota tego typu nie jest możliwe bez wyposażenia go w rodzaj inteligencji emocjonalnej czy społecznej oraz możliwość przeprowadzania konwersacji, a to z kolei oznacza odwołanie się do wspomnianej wcześniej problematyki związanej z przetwarzaniem języka, rozumieniem kontekstu, etc.

Podsumowanie

Pomimo wrażenia, jakie mogą wywierać wspomniane wyżej współczesne projekty AI, żadnego z nich nie można jak na razie chociażby podejrzewać o posiadanie inteligencji zbliżonej do ludzkiej, cały czas mamy do czynienia z *artificial servility*. Z punktu widzenia postronnego obserwatora komputery potrafią obecnie

dokonywać cudów, w rzeczywistości jednak postęp, jaki dokonał się w dziedzinie sztucznej inteligencji od lat 50tych nie jest aż tak duży, jak mogłoby się wydawać (i jak chcieliby futurologzy), a cel, jakim jest stworzenie SSI, nadal wydaje się być odległy.

II. CZY MASZYNY MOGĄ MYŚLEĆ?

“Computational Intelligence is the manifest destiny of computer science, the goal, the destination, the final frontier. More than any other field of science, our computer science concepts and methods are central to the quest to unravel and understand one of the grandest mysteries of our existence, the nature of intelligence. Generations of computer scientists to come must be inspired by the challenges and grand challenges of this great quest.”¹⁶

Edward Feigenbaum

Cele sztucznej inteligencji bywają różnie formułowane. Niektórym wystarcza tworzenie AI działającej racjonalnie. Inni chcieliby widzieć u maszyn przede



Ilustracja II - 1: „Can Machines Think? They already do, say scientists. [...]”

Źródło:<http://content.time.com/time/covers/0,16641,19960325,00.html>

wszystkim racjonalne rozumowanie. Jeszcze inni woleliby, żeby komputery nie tylko zachowywały się racjonalnie, ale także „po ludzku”. Wreszcie, są i tacy, którzy uważają, że celem AI jest stworzenie maszyn nie tylko działających, ale również myślących tak jak człowiek.

Najbardziej ambitne ze wszystkich zadań, stworzenie inteligencji dorównującej ludzkiej, a może nawet ją przewyższającej, jest jednocześnie obarczone największym ryzykiem porażki. SSI może zwyczajnie nie być możliwa i to niezależnie od ilości zasobów

¹⁶ „Sztuczna inteligencja jest „objawionym przeznaczeniem” informatyki, celem, dążeniem, ostateczną granicą. Metody i koncepcje wykorzystywane w informatyce, bardziej niż w jakiegokolwiek innej dziedzinie nauki, są kluczowe dla rozwikłania i zrozumienia jednej z największych tajemnic naszego istnienia, natury inteligencji. Następne pokolenia informatyków muszą być zainspirowane trudnościami i wyzwaniem tego wielkiego przedsięwzięcia.“ [tłumaczenie własne], Edward A. Feigenbaum, *Some Challenges and Grand Challenges for Computational Intelligence*, Journal of the ACM, 2003, nr 1, s. 39

jakimi dysponujemy, czy też jakimi będziemy dysponować w przyszłości.

Niezależnie od hurraoptymizmu niektórych futurologów czy dziennikarzy, pytanie o możliwość stworzenia sztucznej inteligencji na ludzkim poziomie bynajmniej nie sprowadza się do problemu możliwości sprzętowych dzisiejszych komputerów. Kontrowersje dotyczą zarówno kwestii technicznych, jak i filozoficznych, a nawet religijnych.

Wreszcie, zakładając że stworzenie sztucznej inteligencji na ludzkim poziomie jest w ogóle możliwe, pozostaje pytanie, czy byłoby to korzystne dla ludzkości z psychologicznego, socjologicznego, finansowego, a nawet ewolucyjnego punktu widzenia.

Rozważania na temat filozofii sztucznej inteligencji mogą wydawać się niepotrzebnie rozwlekłe i pozbawione praktycznej wartości, a niektóre koncepcje dotyczące przyszłości komputeryzacji naiwne lub wręcz zakrawające o science fiction. Co gorsza, poszukiwania odpowiedzi na tytułowe pytanie mogą okazać się bezowocne i być może temat ten nigdy nie zostanie definitywnie zamknięty. Pomimo wszystko (cytuując fragment wykładu profesora Geoffrey'a Jeffersona): „*I am encouraged [...] to proceed in the hope that, although we shall not arrive at certainty, we may discover some illumination on the way.*”¹⁷

¹⁷ „Zamierzam kontynuować w nadziei, że mimo, że możemy nie osiągnąć pewności, to być może oświecenie spotka nas po drodze.”, [tłumaczenie własne], Cytat pochodzi z wykładu opublikowanego w British Medical Journal, w 1949 roku pt. „The Mind of Mechanical Man”, a wygłoszonego w Royal College of Surgeons of England 9 czerwca 1949 roku z okazji otrzymania przez prof. Jeffersona nagrody Listera . Najbardziej znany fragment tego wykładu zostanie przytoczony w dalszej części rozdziału.

II.1. PRAWO MOORE'A I PRAWO KURZWEIL'A

"He uses statistics as a drunken man uses lamp-posts - for support rather than illumination."¹⁸

Cytat przypisywany Andrew Lang'owi

Niektórzy futurologowie upatrują dowodu na możliwość powstania SSI w samej tylko prędkości rozwoju technologii, przeważnie cytując przy tym Prawo Moore'a, lub porównując pojemności pamięci komputerów z pojemnością mózgu. Jednym z barwniejszych przykładów są tutaj analizy Ray'a Kurzweil'a:

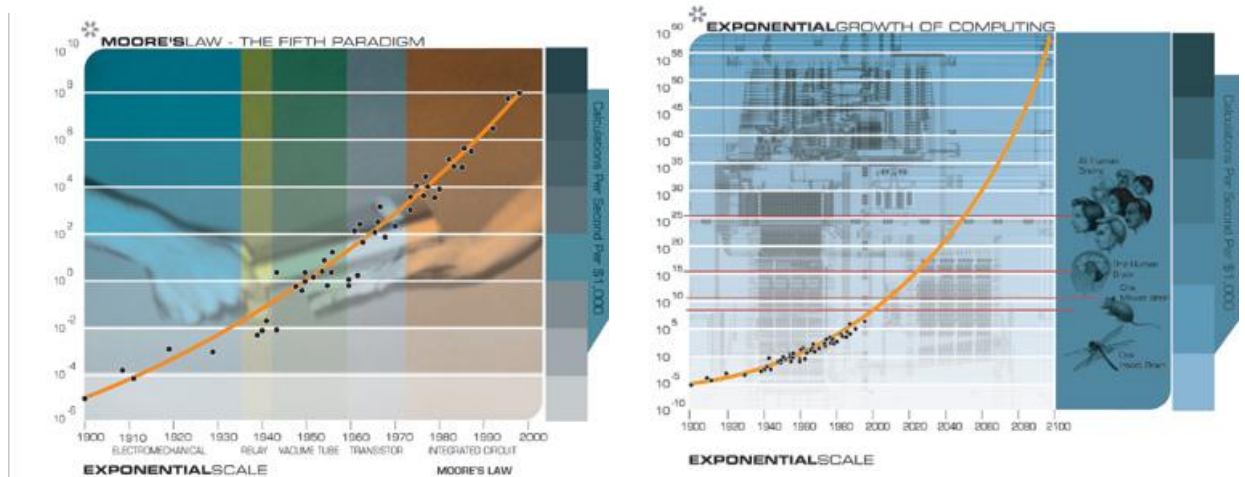
„An analysis of the history of technology shows that technological change is exponential, contrary to the common-sense “intuitive linear” view. So we won't experience 100 years of progress in the 21st century — it will be more like 20,000 years of progress (at today's rate).” (“Analiza historii technologii pokazuje, że jej rozwój ma charakter wykładniczy, w przeciwieństwie do zdroworozsądkowego »intuicyjnie linearnego« podejścia. Tak więc w XXI wieku nie będziemy świadkami 100 lat postępu – to będzie raczej 20.000 lat (zakładając utrzymanie dzisiejszego tempa)”)¹⁹

Kurzweil nie ogranicza się do samego prawa Moore'a, ale proponuje jego rozszerzoną wersję. Gordon Moore, na podstawie analizy historii rozwoju komputeryzacji, stwierdził, że liczba tranzystorów w układach scalonych podwaja się co dwa lata. Teza, że tempo rozwoju technologii jest wykładnicze, nie jest jednak dla Ray'a Kurzweila wystarczająca. Uważa on, że tempo rozwoju samego tempa rozwoju (!) technologii również jest wykładnicze. Twierdzenie to nosi nazwę Prawa Kurzweil'a albo Prawa o przyspieszającym zwrocie (Law of Accelerating Returns).

¹⁸ “Używa statystyki jak pijany człowiek używa latarni – raczej dla wsparcia niż dla oświecenia”, [tłumaczenie własne]

¹⁹ [tłumaczenie własne], Ray Kurzweil, *The Law of Accelerating Returns*, [online]. [dostęp: 15.10.2013]. Dostępny: <http://www.kurzweilai.net/the-law-of-accelerating-returns>

The "returns," such as chip speed and cost-effectiveness, also increase exponentially. There's even exponential growth in the rate of exponential growth. Within a few decades, machine intelligence will surpass human intelligence [...]. The implications include the merger of biological and nonbiological intelligence, immortal software-based humans, and ultra-high levels of intelligence that expand outward in the universe at the speed of light." ("»Zwrot«, taki jak prędkość czipów i opłacalność produkcji również rośnie wykładniczo. Nawet tempo wzrostu tempa wykładniczego wzrostu jest wykładnicze. W przeciągu kilku dekad maszynowa inteligencja przeskoczy ludzką [...]. Do konsekwencji należy fuzja biologicznej i niebiologicznej inteligencji, nieśmiertelni, oparci na oprogramowaniu, ludzie, i ultra-wysoki poziomy inteligencji, który rozprzestrzenią się we wszechświecie z prędkością światła.")²⁰



Ilustracja II - 2 Wykresy przedstawiające odpowiednio: prawo Moore'a i prawo Kurzweila na skali logarytmicznej. Źródło: Ray Kurzweil, The Law of Accelerating Returns

Co ciekawe, sam Moore uważa, że tempo przyspieszania działania układów scalonych w najbliższym czasie ulegnie raczej zmniejszeniu (około dwukrotnemu) i zdecydowanie nie wierzy w Osobliwość Technologiczną. Nie jest zresztą jedynym naukowcem podchodzącym sceptycznie do tez głoszonych przez autora „The Singularity Is Near”. Krytycy oskarżają Kurzweil’a o uprawianie paranauki i szerzenie eskapistycznych fantazji. Popularność „singularianizmu” tłumaczą quasi religijnym charakterem teorii o Osobliwości, potęgowanym przez prognozy Kurzweila dotyczące nieśmiertelności.

²⁰ Tamże

Do innych entuzjastów stanowiska opartego na mocy obliczeniowej należą Bill Joy i Hans Moravec.

W gąszczu statystyk i wyliczeń łatwo przeoczyć podstawowy błąd będący bolączką części rozważań tego typu. Otóż entuzjaści stanowiska o zbliżającej się nieuchronnie Osobliwości, czy też po prostu inteligencji dorównującej ludzkiej, odpowiadają wyłącznie na pytanie *kiedy* maszyny zaczną myśleć. Jednocześnie pomijają zupełnie kwestię tego, *czy* maszyny w ogóle są w stanie myśleć. Nie ma przekonujących dowodów na to, że przyspieszenie działania komputerów wiąże się w jakikolwiek sposób z przyspieszeniem komputerowej myśli, bo nie wiemy jak taka myśl miałaby działać. Są też poważne wątpliwości co do tego, czy maszyny analogiczne do maszyny Turinga są właściwymi narzędziami do zadania stworzenia SSI (o czym dalej).

II.2. ZARZUTY, SPRZECIWIWY, ARGUMENTY...

II.2.1. COMPUTING MACHINERY AND INTELLIGENCE

Specjalistów w dziedzinie sztucznej inteligencji od lat trapią problemy natury filozoficznej, logicznej, obliczeniowej i moralnej. Wiele z tych problemów zostało zasygnalizowanych już ponad 60 lat temu, w czasach EDSAC'a i Mark'a I.

Alan Turing jest uważany za jednego z ojców współczesnej komputeryzacji. W słynnym artykule z 1950 roku pt. *„Computing Machinery and Intelligence”*, przedstawił szereg prawdopodobnych argumentów przeciwko możliwości powstania komputerowej myśli. Każdy z nich rozpatrzył i (w mniejszym lub większym stopniu) odparł. Jest to o tyle ciekawe, że pomimo upływu czasu i niewyobrażalnego rozwoju technologii, dyskusje o AI często nadal krążą wokół przewidzianych przez Turinga osi sporów.

Przed przejściem do obiekcji przedstawianych przez Turinga nie można pominąć jego własnej odpowiedzi na pytanie „Czy maszyny mogą myśleć”. A właściwie jej braku.

“The original question, ‘Can machines think?’ I believe to be too meaningless to deserve discussion.” (“Oryginalne pytanie , ‘Czy maszyny mogą myśleć’ uważam za zbyt pozbawione znaczenia by zasługiwało na dyskusję”²¹)

Turing uważał, że problemy związane ze sceptycznym nastawieniem ludzi do mechanicznej myśli wynikają po części z niedostatecznie precyzyjnego (czy też nie dość odpowiadającego zmianom technologicznym) słownictwa. Wierzył, że do końca XX wieku znaczenie słów zdąży na tyle wyewoluować, że mówienie o myślących maszynach będzie zupełnie naturalne (*„[...] the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.”*²²). Uznał, że odpowiedź na pytanie „czy maszyny mogą myśleć?” wymaga koniecznie zdefiniowania słów „maszyna” oraz „myśleć”. Jednocześnie zdecydowanie odrzucił pomysł określenia

²¹ [tłumaczenie własne], Alan Turing, *Computing Machinery and Intelligence* [online]. [dostęp: 5.10.2013]. Dostępny: <http://www.turingarchive.org/browse.php/B/9>

²² *Tamże*

ich znaczenia w oparciu o statystycznie najczęstsze sposoby ich używania, a w zamian za to zaproponował słynny test behawioralny (Test Turinga), który zostanie szerzej omówiony w rozdziale trzecim.

Podobne stanowisko możemy znaleźć u Marviniego Minsky'ego (wpominanego już w poprzednim rozdziale kognitywisty i jednego z pionierów sztucznej inteligencji):

„Saying that someone is like a machine has come to have two opposite meanings: (1) „To have no intentions, goals, or emotions,” and (2) “to be relentlessly committed to a single purpose or policy.” Each meaning suggest inhumanity, as well as a kind of stupidity because excessive commitment results in rigidity, while lack of purpose leads to aimlessness. However, [...] both of those views will be obsolete [...].” (“Przyjęło się uważać, że stwierdzenie, że ktoś jest jak maszyna, ma dwa przeciwstawne znaczenia (1) »Nie posiadać intencji, celów i emocji«, i (2) »być bezwzględnie oddanym jednemu celowi lub jednej polityce«. Oba znaczenia sugerują nieludzkość i rodzaj głupoty, ponieważ nadmierne zaangażowanie skutkuje zeszywnieniem, a brak celu prowadzi do bezcelowości. Jednak [...] oba te poglądy staną się nieaktualne”)²³

Oraz u Dijkstry:

“The question of whether Machines Can Think... is about as relevant as the question of whether Submarines Can Swim.” (“Pytanie czy Maszyny Mogą Myśleć... jest mniej więcej tak samo sensowne jak pytanie czy Łodzie Podwodne Mogą Pływać.”)²⁴

Zakładając jednak tymczasowo, że tytułowe pytanie nie jest całkiem bezzasadne, czas powrócić do argumentów proponowanych przez Turinga.

²³ Marvin Lee Minsky, *The Emotion Machine. Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Nowy York: SIMON&SCHUSTER PAPERBACKS, 2006, s. 33 [tłumaczenie własne]

²⁴ (transkrypt konferencji) E.W. Dijkstra, *The threats to computing science - ACM South Central Regional Conference* [online]. [dostęp: 14.10.2013]. Dostępny: <http://www.cs.utexas.edu/~EWD/transcriptions/EWD08xx/EWD898.html> [tłumaczenie własne]

II.2.1.1. Sprzeciw teologiczny:²⁵

A: Myślenie jest funkcją duszy człowieka. Z kolei dusza jest dana ludziom przez Boga, dlatego też istnienie myślących maszyn jest niemożliwe.

O: Turing nie uznaje argumentów teologicznych za szczególnie znaczące. Mimo to podejmuje sprytną próbę zbitcia powyższej obiekcji, stwierdzając, że teza w niej zawarta wydaje się nakładać ograniczenia na boską moc. Argumentuje też, że tworzenie myślących maszyn jest uzurpowaniem sobie mocy Boga w nie większym stopniu niż poczynanie potomstwa.

Mimo, że argument ten nie jest raczej poważnie rozpatrywany przez uczonych, to różne jego odmiany są nadal żywe w dyskusjach laików. Przykładowo, ciekawe wariacje na temat odpowiedzi Turinga można znaleźć na katolickich forach internetowych:

“Once a thinking being exhibits every sign of humanity... intelligence, emotion, free-will, etc.... it seems that from a practical sense, our only option is to assume it has a soul and accord it all the rights of any human being. The right to pray. The right to baptism, the sacrament, and confession? The right to become a priest? We would have absolutely no evidence or grounds for denying this creature those rights and would be compelled to leave that up to God.” (“W momencie, w którym myślące stworzenie będzie wykazywało wszystkie znaki bycia człowiekiem... inteligencję, emocje, wolną wolę, etc... wydaje się, że z praktycznego punktu widzenia, naszą jedyną opcją jest założyć, że ma ono duszę i przypisać mu wszystkie prawa przynależne każdej istocie ludzkiej. Prawo do modlitwy. Prawo do chrztu, do sakramentu, do spowiedzi. Prawo zostania księdzem? Nie mielibyśmy absolutnie żadnego dowodu ani żadnej podstawy żeby odmówić takiemu stworzeniu tych praw i bylibyśmy zobowiązani zostawić to w gestii Boga”)²⁶

²⁵ Obiekcje przedstawiam w formie argumentów („A:”) i odpowiedzi („O:”) Turinga; w podtytułach zachowuję oryginalne nazewnictwo Turinga

²⁶ Catholic Answers [online]. [dostęp: 13.10.2013]. Wypowiedź użytkownika Scott Walker. Dostępny: <http://forums.catholic.com/showthread.php?t=106016&page=15>, [tłumaczenie własne]

II.2.1.2. Argument „głów w piasku” ('Heads in the Sand' Objection):

A: Nie należy tworzyć myślących maszyn, ponieważ będzie to miało straszliwe konsekwencje.

Q: W tym przypadku Turing stwierdza, że argumenty tego typu nie są wystarczająco namacalne, żeby się do nich poważnie odnosić.

Bardziej sprecyzowanym i „namacalnym” argumentom tego rodzaju poświęcony jest podrozdział II.3.

II.2.1.3. Argument matematyczny:

A: W jakimkolwiek odpowiednio rozwiniętym systemie logicznym, można sformułować twierdzenie, którego prawdziwości albo nieprawdziwości nie będzie można stwierdzić w obrębie tego systemu.

Q: Odpowiedzią Turinga jest stwierdzenie, że nie ma dowodu na to, że ludzki umysł nie jest obwarowany podobnymi ograniczeniami.

Argument ten często nazywany jest argumentem z twierdzenia Gödel'a. Twierdzenie to wywołało zdumienie najwybitniejszych matematyków podczas konferencji w Królewcu w 1930 roku. Gödel wykazał, że „żaden formalny system poprawnych reguł dowodzenia w matematyce nie wystarcza, nawet „w zasadzie”, aby udowodnić wszystkie prawdziwe twierdzenia zwyczajnej arytmetyki”. Sir Roger Penrose, z którego książki pt. „Cienie Umysłu”(1994) pochodzą cytaty w tym akapicie, czyni to stwierdzenie podstawą swojej tezy, że „ludzkie myślenie wymaga czegoś, co jest nieosiągalne dla komputerów, w takim sensie, w jakim rozumiemy pojęcie »komputer« obecnie. W pośredniej konsekwencji Penrose szuka źródeł dla zjawiska świadomości w „niealgorytmicznych procesach fizycznych”.

II.2.1.4. Argument ze świadomości:

A: Dopóki maszyna nie będzie w stanie czuć, że myśli, nie będzie naprawdę myśląca. Turing jako ilustrację tego argumentu cytuje często przytaczany w literaturze fragment wspomnianego już wykładu profesora Geoffrey'a Jeffersona, który przytaczam poniżej:

„Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and no: by the chance fall of symbols, could we agree that machine equals brain - that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry, or depressed when it cannot get what it wants.” („Dopóki maszyna nie będzie potrafiła napisać sonetu lub skomponować koncertu powodowana myślami i przeżyтыми emocjami, a nie dzięki przypadkowemu złożeniu symboli, nie będziemy mogli się zgodzić, że maszyna jest równa mózgowi – czyli, nie tylko to pisze, ale wie, że to napisała. Żaden mechanizm nie może czuć (a nie tylko sztucznie sygnalizować czucie, przy pomocy prostych sztuczek) przyjemności z sukcesu, smutku kiedy przepalają mu się zawory, być rozgrzanym przez komplement, przygnębionym przez własne błędy, oczarowanym płcią, złym lub zdołowanym kiedy nie może dostać tego, czego chce.”)²⁷

Q: Turing odpowiada, że nie jesteśmy w stanie z pewnością stwierdzić, że inni ludzie myślą i czują tak jak my. Zamiast próbować to sprawdzać, uprzejmie przyjmujemy („polite convention”), że otaczające nas osoby myślą, na podstawie ich zachowania oraz ich podobieństwa do nas samych. W związku z tym pozostaje nam albo porzucenie tego argumentu, albo przyjęcie radykalnej, solipsystycznej²⁸, postawy.

Temat świadomości inteligentnych maszyn należy do najbardziej kontrowersyjnych, ze względu na to, że dotyczy czegoś, czego każdy z nas

²⁷ Geoffrey Jefferson, *The Mind of Mechanical Man*, British Medical Journal, 1949, t. 1, nr 4616, s.1110 [tłumaczenie własne]

²⁸ Solipsyzm jest poglądem filozoficznym, zgodnie z którym poznawalna jest tylko własna świadomość podmiotu poznającego.

bezpośrednio i subiektywnie doświadczą. W pewnym sensie każdy z nas jest ekspertem w sprawie własnej świadomości. Turing nie jest szczególnie zainteresowany dowodzeniem możliwości lub niemożliwości istnienia świadomych maszyn.

Teoretycznie temat świadomości można spróbować ominąć, stwierdzając, że problematyka możliwości powstania *świadomej* myśli wykracza poza podstawowe pytanie zadane w tytule. Sądzę jednak, że sztuczna inteligencja na ludzkim poziomie jest często utożsamiana z samoświadomością maszyn, w związku z czym pytanie o to czy maszyny mogą myśleć może być rozumiane jako pytanie o to czy maszyny mogą myśleć *świadomie*. Próby rozwiązania zagadki świadomości prowadzą też do lepszego zrozumienia działania ludzkiego umysłu, a przez to, w konsekwencji, służą poprawie działania programów sztucznej inteligencji. Temat ten będzie rozwinięty w podrozdziale II.2.2.3..

II.2.1.5. Argumenty z różnych niemożności (Arguments from Various Disabilities):

A: Mowa tu o argumentach postaci: „Maszyna nie jest zdolna do/nie jest w stanie zrobić X”. Turing podaje wiele przykładów, np. maszyna nie jest w stanie być miłą/piękną, zakochać się, czy mieć przyjemności z jedzenia truskawek ze śmietaną, etc.

Q: Turing uznaje, że do większości tego typu argumentów nie trzeba się specjalnie odnosić, bo jest tylko kwestią czasu, kiedy przestaną być aktualne. Pozostałe redukuje do argumentu ze świadomości.

Jak widać na podstawie przeglądu zawartego w rozdziale pierwszym, rzeczywiście niektóre z tego typu argumentów się zdezaktualizowały. Maszyny pokonują ludzi w grach, zastępują ekspertów, uczą się na własnych błędach. Jednak wiele z dawnych obiekcji nadal pozostaje w mocy, szczególnie tych dotyczących emocjonalnej sfery naszego funkcjonowania.

II.2.1.6. Sprzeciw Lady Lovelace

A: Lady Ada Lovelace (właściwie Augusta Ada Byron, córka Lorda Byrona), uznawana za pierwszego programistę w historii (urodzona w 1815 roku), pisała programy dla nigdy nieukończonych Maszyny Analizacyjnej Babbage'a. Jest też autorką spekulacji na temat prawdopodobnych możliwości maszyn.

„The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform” (“Maszyna Analizacyjna nie rości sobie pretensji by dawać czemukolwiek początek. Może zrobić cokolwiek umiemy rozkazać jej wykonać”)²⁹

Argument ten możemy streścić jako „Maszyna nigdy nie zrobi niczego nowego”. Turing jako ulepszoną wersję tego argumentu proponuje „Maszyna nigdy nas nie zaskoczy”.

O: Turing odpowiada, w nieco żartobliwym tonie, że jego samego maszyny często zaskakują. Pytanie czy maszyny mogą wprawić nas w zdumienie częściowo sprowadza (po raz kolejny) do problemu świadomości, a częściowo do (błédnego, zdaniem Turinga) założenia, że maszyna działająca według swoich pierwotnych założeń nie może wytworzyć niczego zaskakującego.

Sprzeciw Lady Lovelace jest właściwie szczególnym przypadkiem argumentu z różnych niemożności. Być może jego wyodrębnienie przez Turinga wynikało z historycznego znaczenia i związanego z nim częstego występowania tej obiekcji w dyskusjach dotyczących AI.

II.2.1.7. Argument z ciągłości układu nerwowego:

A: Układ nerwowy, w przeciwieństwie do komputerów, nie funkcjonuje w sposób dyskretny.

²⁹ Ada Augusta Byron, notatki do *Sketch of The Analytical Engine Invented by Charles Babbage L. F. Menabrei* [online]. [dostęp: 15.10.2013]. Dostępny: <http://psychclassics.yorku.ca/Lovelace/lovelace.htm> [tłumaczenie własne]

Q: Turing odpowiada, że być może maszyna działająca w sposób dyskretny nie będzie w stanie zduplikować działania maszyny nie posiadającej takiego ograniczenia, ale za to może ją symulować.

Argument ten występuje nadal, chociaż w nieco zmienionej formie. Dość znany jest pogląd, że wszystko, co jest związane z możliwościami poznawczymi ludzkiego umysłu może być zduplikowane przez maszynę Turinga. W takim ujęciu umysł jest traktowany jako mechanizm przetwarzający informacje i jako taki da się zamodelować przy pomocy UTM (uniwersalnej maszyny Turinga)³⁰

Tak więc, maszyna będzie mogła działać tak jak człowiek jeżeli ludzki umysł jest redukowalny do maszyny Turinga (stanowisko to nosi nazwę „narrow mechanist view”, przeciwny pogląd - „wide mechanist view” - opiera się na analogicznym założeniu, że system przetwarzania informacji nie daje się sformalizować jako maszyna Turinga)³¹.

Można więc roboczo zastąpić pierwotny argument pytaniem o to czy maszynę, która nie funkcjonuje w sposób dyskretny, da się zredukować do maszyny Turinga. Odpowiedź na tak postawione pytanie jest znana - niektóre maszyny analogowe są nieredukowalne do maszyny Turinga (ze względu na to, że nie wszystkie liczby rzeczywiste są obliczalne przy pomocy maszyny Turinga).

Jednym z przykładów maszyn niesprowadzalnych do maszyny Turinga jest ETM (Extended Turing Machine). ETM w każdym kwadracie taśmy przechowuje dowolną liczbę rzeczywistą (obliczalną lub nie), natomiast maszyna Turinga przechowuje tylko pojedyncze symbole dyskretny. W związku z tym ETM nie daje się sprowadzić do UTM.

Skoro maszyna działająca w sposób dyskretny nie będzie mogła działać tak jak ludzki umysł, to może, zgodnie z sugestią Turinga, będzie mogła go symulować?

³⁰ UTM, Uniwersalna Maszyna Turinga (Universal Turing Machine), maszyna Turinga posiadająca zdolność symulowania dowolnej innej maszyny Turinga

³¹ Nazewnictwo podają za : The Oxford Handbook of Philosophy of Cognitive Science, Oxford University Press

Taka symulacja jest możliwa zgodnie z założeniem, że każda maszyna o skończonej ilości komunikatów wyjściowych (a człowiek kiedyś umiera, więc jego liczba komunikatów jest skończona) jest symulowalna przez maszynę Turinga. Mowa tu jednak jedynie o teoretycznej możliwości symulowania umysłu. W praktyce mogłoby okazać się, że taka symulacja wymagałaby np. istnienia sygnałów poruszających się z prędkością przekraczającą prędkość światła.³²

II.2.1.8. Argument z nieformalności zachowania:

A: Nie istnieje skończony zestaw zasad postępowania regulujących ludzkie życie (innymi słowy: ludzkie zachowanie jest zbyt skomplikowane by dało się je ująć przy pomocy prostego zbioru reguł).

O: Turing odpowiada, że być może faktycznie nie istnieje skończony zestaw reguł postępowania, ale nie możemy wykluczyć takiej możliwości. To, że nie udało się znaleźć kompletnego zbioru reguł nie jest jeszcze dowodem na to, że taki zbiór nie ma prawa istnieć. Używając sformułowania przypisywanego Martinowi Rees'owi³³ - nieistnienie dowodu nie jest dowodem nieistnienia.

Argument ten należy do najważniejszych w krytyce sztucznej inteligencji i jest szerzej znany pod nazwą problemu kwalifikacji (qualification problem). Do głównych zwolenników tego argumentu należą bracia Dreyfus. Aby lepiej wytłumaczyć skąd wziął się problem kwalifikacji, trzeba sięgnąć do historii sztucznej inteligencji, a konkretnie do pojęcia „Starej, Dobrej Sztucznej Inteligencji” („Good Old-Fashioned AI”, w skrócie GOFAI). Termin ten został wymyślony przez Johna Haugelanda i odnosi się do pierwotnego paradygmatu AI.

Wczesne próby stworzenia myślących maszyn zaowocowały hipotezą, że każdy inteligentny system musi operować (na zasadzie logicznego wnioskowania) na symbolicznych strukturach danych. Krytyka Huberta

³² Przykład za: *The Oxford Handbook of Philosophy of Cognitive Science*, Oxford University Press

³³ Postać Sir Martina Rees'a zostanie dokładniej omówiona w rozdziale III.

Dreyfusa oparta na problemie kwalifikacji odnosi się właśnie do takiego podejścia do sztucznej inteligencji. Teoretycznie więc, argument z nieformalności zachowania nie odnosi się do całej dziedziny sztucznej inteligencji, ale raczej do jednego, konkretnego podejścia do jej tworzenia. W praktyce jednak problem ten nie jest całkowicie rozwiązany (częściową odpowiedzią na problem kwalifikacji jest logika niemonotoniczna).

II.2.1.9. Argument z percepcji pozazmysłowej:

A: Właściwie nie jest to argument przeciwko sztucznej inteligencji w ogóle, a raczej problem, który mógłby zaburzyć proces sprawdzania, czy maszyna jest inteligentna. Pod pojęciem percepcji pozazmysłowej Turing rozumie telekinezę, jasnowidzenie i prekognicję.

Q: Turing wydaje się traktować ten problem całkiem serio i proponuje, w razie gdyby doniesienia o fenomenach pozazmysłowych miały się potwierdzić, zastosowanie odpowiednich środków zaradczych („telepathy proof room”, pokój odporny na telepatię) w trakcie testów.

Argument ten ma w tej chwili raczej wartość historyczną. Gdyby chciał go faktycznie dzisiaj rozpatrywać, to prawdopodobnie byłyby on częścią argumentacji dualistycznej, zakładającej rozłączność ducha i materii.

Podsumowanie:

Jak widać, kolejność, w jakiej Turing przytacza argumenty może wydawać się nieco przypadkowa. Część z nich jest argumentami przeciwko słabej sztucznej inteligencji (II.2.1.3., II.2.1.8.), część – silnej (II.2.1.1., II.2.1.2., II.2.1.4., II.2.1.5., II.2.1.6. , II.2.1.7.), ostatni jest już właściwie tylko argumentem przeciwko samej formule Testu Turinga. Trzeba jednak pamiętać, że test jest tutaj nadrzędnym tematem dla Turinga, a koncepcje słabej i silnej sztucznej inteligencji nie miały jeszcze nawet szansy powstać.

Turing, w *Computing Machinery and Intelligence*, pisząc „computer” miał na myśli „human computer”³⁴, a więc człowieka wyposażonego w długopis i zapas papieru. Mimo to przewidział wiele z problemów, które trapią sztuczną inteligencję do dziś. Na przykładzie argumentów Turinga można też prześledzić jak rozwijała się myśl towarzysząca rozwojowi technologii.

II.2.2. INNE ARGUMENTY

II.2.2.1. Problem ramy odniesienia (frame problem):

Problem ramy odniesienia jest kolejnym (po problemie kwalifikacji i argumencie matematycznym) który dotyczy nie tylko silnej, ale i słabej sztucznej inteligencji. Pierwszą intuicją dotyczącą projektowania inteligentnych (w słabym lub silnym sensie) agentów może być wyposażenie ich w jak największą ilość wiedzy. Duża ilość faktów może się jednak okazać dla maszyny (programu, agenta) bardziej przeszkodą niż ułatwieniem, jeżeli nie wprogramujemy mu jednocześnie jakiegoś mechanizmu filtrowania zapamiętanych danych i nowych informacji przychodzących z zewnątrz w odniesieniu do zastanej sytuacji.

„A walking encyclopedia will walk over a cliff, for all its knowledge of cliffs and the effects of gravity, unless it is designed in such a fashion that it can find the right bits of knowledge at the right time, so it can plan its engagement with the real world.” (“Chodząca encyklopedia spadnie z klifu, pomimo całej swojej wiedzy o klifach i efektach działania grawitacji, chyba że zostanie zaprojektowana w taki sposób, że będzie potrafiła znaleźć właściwe informacje we właściwym czasie, żeby móc zaplanować swoją interakcję ze światem rzeczywistym.”)³⁵

Rozważenie wszystkich konsekwencji danej akcji może się wydawać czasochłonne lub trudne. Prawdziwy problem zaczyna się jednak wtedy,

³⁴ „computer” może oznaczać po angielsku zarówno komputer, jak i kalkulator (osobę wykonującą kalkulacje)

³⁵ Daniel Dennett, *Cognitive Wheels: The Frame Problem of AI*. W: Minds, Machines, and Evolution. Red. C. Hookway. Cambridge University Press, 1984, s. 128-151

kiedy musimy rozpatrzyć wszystkie te rzeczy, które się nie ulegną zmianie. Problem ramy odniesienia sprowadza się właśnie do tego – jak przedstawić efekt akcji, ale jednocześnie nie zajmować się wszystkim tym, co nie będzie mieć miejsca.

W latach osiemdziesiątych problem ramy odniesienia był przez niektórych uważany za gwóźdź do trumny projektu stworzenia sztucznej inteligencji. Dotychczas pozostaje on do końca nierozwiązany. W tym przypadku jedynie czas może pokazać czy frame problem może być ostatecznym argumentem przeciwko AI.

II.2.2.2. Chiński pokój:

Argument chińskiego pokoju, autorstwa Johna Searle'a jest jednym z najbardziej znanych argumentów przeciwko istnieniu (silnej) sztucznej inteligencji. Searle proponuje następującą hipotetyczną sytuację: mamy urzędnika mówiącego wyłącznie po angielsku („clerk”), który siedzi zamknięty w pokoju, wyposażonym w książki z instrukcjami (a właściwie z programem), ołówek, papier i dwa otwory (input i output). Zadaniem urzędnika jest przetworzenie podawanych mu arkuszy w języku chińskim według instrukcji zawartej w książkach i zwrócenie ich na zewnątrz. Wykonuje on wszystkie zadania zawarte w instrukcjach, tak jak wykonałby je program komputerowy.

Przekazywane urzędnikowi arkusze zawierają historię po chińsku i pytania do tej historii, natomiast książki zawierają algorytmy pozwalające na mozolne wygenerowanie odpowiedzi na te pytania (również po chińsku). Przebywający w „chińskim pokoju” człowiek nie ma jednak o tym wszystkim pojęcia. On sam jedynie wykonuje instrukcje przetwarzając nieznanne mu symbole. Dla osoby z zewnątrz „pokój” zachowuje się jednak jak osoba mówiąca po chińsku. Mimo to urzędnik wewnątrz nadal nie rozumie w tym języku ani słowa. Searle twierdzi, że z tego właśnie powodu komputer również nie może niczego rozumieć.

Przyjęcie, że urzędnik faktycznie nie rozumie języka chińskiego nie wyklucza jednak, że nie rozumie go też sam „pokój”. Pokój jest w tym eksperymencie myślowym odpowiednikiem komputera, więc tak naprawdę kontrargument sprowadza się do tego, że fakt, że jakaś część systemu nie rozumie chińskiego, nie oznacza że cały system również go nie pojmuje.

II.2.2.3. Świadomość, zarys wstępu do podstaw, czyli argument ze świadomości raz jeszcze:

Wczoraj i dziś:

John Searle, w swoim ostatnim wystąpieniu w ramach TEDxCERN³⁶ zastanawiał się dlaczego temat świadomości nie jest bardziej obecny w nauce i kulturze.

“Now why is that curious? Well. it is the most important aspect of our lives for a very simple, logical reason, namely, it’s a necessary condition on anything being important in our lives that we’re conscious. You care about science, philosophy, music, art, whatever -- it’s no good if you’re a zombie or in a coma, right? So consciousness is number one.” (“Dlaczego to jest zastanawiające? Cóż, jest to najważniejszy aspekt naszego życia z bardzo prostego, logicznego powodu, mianowicie, koniecznym warunkiem, żeby cokolwiek w naszym życiu było istotne, jest to, żebyśmy byli świadomi. Interesuje Cię nauka, filozofia, muzyka, sztuka, cokolwiek – kiepsko jeżeli jesteś zombie albo w śpiączce, prawda? Więc świadomość jest numerem jeden.”)³⁷

Searle upatruje przyczyny niewielkiego zainteresowania tematem świadomości w spadku po behawioryzmie.

³⁶ TED (Technology, Entertainment and Design) – konferencja organizowana przez Sapling Foundation, odbywa się od 1984 roku, od 2006 roku wystąpienia w ramach TED udostępniane są w Internecie na licencji Creative Commons na stronie www.ted.com. TEDxCERN, o którym mowa, został zorganizowany na licencji TED w ośrodku CERN w maju 2013 roku.

³⁷ Wystąpienie Johna Searle’a w ramach TEDxCERN.[online].[dostęp: 12.10.2013], Dostępny: http://www.ted.com/talks/john_searle_our_shared_condition_consciousness.html, [tłumaczenie własne]

"I said that behaviorism ought to be one of the great embarrassments of our intellectual culture [...]. The mistake was [...] they [behaviorists] think if you accept the irreducible existence of consciousness, you're giving up on science. You're giving up on 300 years of human progress and human hope and the rest of it." („Powiedziałem, że behawioryzm powinien być jednym z największych powodów do wstydu naszej intelektualnej kultury [...]. Błąd polegał na tym [...], że oni [behawioryści] myślą, że akceptacja istnienia nieredukowalnej świadomości oznacza porzucenie nauki. Porzucenie 300 lat postępu i ludzkiej nadziei i całej reszty”)

Faktycznie, rozwój behawioryzmu na początku XX wieku spowodował zahamowanie badań nad świadomością, od lat osiemdziesiątych możemy jednak zaobserwować renesans zainteresowania tym tematem (m.in. w pracach wspomnianych już Sir Rogera Penrose'a i Daniela Dennetta). Zagadka świadomości pozostaje jednak, pod wieloma względami, tak samo tajemnicza jak była np. ponad sto lat temu dla T. H. Huxley'a³⁸:

"How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin, when Aladdin rubbed his lamp." („Jak to możliwe, że coś tak niezwykłego jak stan świadomości jest rezultatem podrażnienia tkanki nerwowej, jest to tak samo niewytłumaczalne jak pojawienie się Dżina po potarciu lampy przez Alladyna. ”)³⁹

Explanatory gap:

To co zadziwiało Huxley'a nosi nazwę "explanatory gap" i jest jednym z głównych problemów badań nad świadomością. W skrócie jest to nieumiejętność znalezienia łącznika między działaniem neuronów a świadomymi procesami myślowymi. Niektórzy filozofowie twierdzą, że nieudane zmagania ludzkości z tym tematem dowodzą tego, że nie jesteśmy w stanie zrozumieć własnej świadomości. Inni z kolei twierdzą, że świadomość nie istnieje, jest tylko iluzją, złudzeniem.

³⁸ Angielski biolog, dziadek pisarza Aldousa Huxley'a

³⁹ T.H.Huxley, *Lessons on Elementary Physiology 8*, [tłumaczenie własne]

Searle odrzuca obie powyższe tezy. Nie zgadza się też z dualistami i materialistami, którzy prezentując radykalnie odmienne od siebie stanowiska, zgadzają się jednak co do tego, że świadomość nie może być częścią światopoglądu naukowego. Uważa, że świadomość poddaje się naukowemu rozumowaniu, nawet pomimo tego, że jest ze swojej natury odczuciem subiektywnym. Argumentuje, że skoro można próbować leczyć pacjentów, którzy doznają subiektywnego odczucia bólu, przy pomocy obiektywnych metod, to ogólnie, można wyciągać obiektywne wnioski o czymś co ma charakter subiektywny. Twierdzi, że świadome stany są wynikiem neurobiologicznych procesów w mózgu.⁴⁰ Nie odpowiada jednak na kluczowe pytanie w jaki sposób jest to możliwe.

Qualia:

Kolejnym, po explanatory gap, tematem wzbudzającym kontrowersje jest istnienie qualiów. Qualia są „surowymi odczuciami”. Według niektórych dopiero posiadanie qualiów czyni jednostkę prawdziwie świadomą.

Teoria o istnieniu qualiów stoi w opozycji do funkcjonalizmu i fizykalizmu (według którego wszystkie poprawne informacje są informacjami fizycznymi). Frank Jackson, który określa siebie samym mianem „qualia freak”, tłumaczy to następująco:

„Tell me everything physical there is to tell about what is going on in a living brain, the kind of states, their functional role, their relation to what goes on at other times in all other brains, and so on and so forth, and be as clever as can be in fitting it all together, you won't have told me about the hurtfulness of pains [...], or about the characteristic experience of tasting a lemon [...].” („Powiedz mi wszystko co można powiedzieć o tym co dzieje się w żywym mózgu, o rodzajach stanów, ich roli funkcjonalnej, ich relacji do tego co dzieje się w innym czasie we wszystkich innych mózgach i tak dalej i tym podobne. Bądź tak zmyślny jak chcesz w dopasowywaniu tego wszystkiego w jedną całość, a nadal nie powiedziałaś mi o bolesności bólu

⁴⁰ W ujęciu Searle'a, świadomość jest realna, nieredukowalna i ma wpływ na rzeczywistość.

[...], czy charakterystycznym odczuciu towarzyszącemu jedzeniu cytryny [...].)”⁴¹

Tak więc, qualia istnieją jeśli fizykalizm coś pomija. Jackson przytacza kilka dowodów na ich istnienie. Jeden z nich to zmyślona historia Freda – człowieka który rozróżnia o jeden odcień czerwieni więcej niż wszyscy inni. Fred mógłby dowolnie długo i dokładnie tłumaczyć nam wszystko na temat różnic między widzianymi przez niego wersjami czerwieni, przekazując nam wszelkie naukowe dane. Jednak mimo to, dopiero gdyby udało się przeszczepić wzrok Freda innym, ludzie cieszyliby się, że wiedzą jak to jest być Fredem.

Temat qualiów jest tym trudniejszy, że z definicji nie mają one wpływu na świat fizyczny. Jak więc wytłumaczyć ich istnienie, jak mogły wyewoluować skoro nie mogły mieć wpływu na przetrwanie jednostki? Jackson tłumaczy, że powstanie qualiów mogło być produktem ubocznym ewolucji, tak jak fakt, że futro niedźwiedzia polarnego jest ciężkie, co samo w sobie nie jest w żaden sposób przystosowawcze, jednak było konieczną konsekwencją tego, że jest ono ciepłe. Natomiast to, że qualia nie mają wpływu na świat fizyczny, nie oznacza, że nie mogą mieć wpływu na inne stany psychiczne (umysłowe).

Pytania, definicje:

Zarówno explanatory gap, jak i qualia są dość obszernymi tematami. Nadal są jednak jedynie wycinkiem badań nad świadomością w ogóle. Znaczna trudność pojawia się już przy próbie jej zdefiniowania. Teoretycznie każdy intuicyjnie wie, co to znaczy być świadomym. W praktyce jednak nie jest to takie proste - Stanford Encyclopedia of Philosophy wymienia 6 różnych kryteriów świadomości (istoty) i tyle samo różnych znaczeń stwierdzenia, że czyjś stan psychiczny (umysłowy) jest świadomy (w tym m.in. że dana osoba posiada qualia).

Nawet jeśli uda się stwierdzić czym jest świadomość, to pozostają jeszcze, nie mniej ważne, pytania – skąd się wzięła, jak to możliwe, że powstała, jak

⁴¹ Frank Jackson, *Epiphenomenal Qualia*, Philosophical Quarterly, 1982, nr 32 [tłumaczenie własne]

działa, dlaczego i po co istnieje. Pomimo funkcjonowania wielu teorii droga do wyjaśnienia zagadki świadomości jest jeszcze daleka. Dlatego też nie można z całą pewnością stwierdzić czy argument ze świadomości dla SSI jest zasadny. Np. jeżeli qualia istnieją i są zjawiskiem wynikającym z niealgorytmicznych procesów fizycznych, to komputery (jakie znamy) prawdopodobnie nie będą w stanie takich odczuć zduplikować. Z kolei jeśli rację mają Ci, którzy twierdzą, że świadomość jest złudzeniem (jak np. Daniel Dennett), to argument ze świadomości traci sens – trudno oskarżać maszynę o brak czegoś, co nie istnieje. Być może świadomość jest tak kluczowym, centralnym konceptem dla funkcjonowania naszych umysłów, że nie da się stworzyć nieświadomej maszyny o inteligencji dorównującej ludzkiej. Nawet jeśli badania nad świadomością nie dadzą nam żadnych jednoznacznych odpowiedzi, to niewykluczone, że przyczynią się do przyspieszenia powstania prawdziwie inteligentnych maszyn, albo też doprowadzą do stworzenia sprytniejszych symulacji.

Podsumowanie:

Drugą część powyższego rozdziału rozpoczyna się od omówienia znanych argumentów, które nie pojawiły się w tekście Turinga, a kończy rozwinięciem tematu świadomości, który zdecydowałam się dokładniej omówić osobno, ze względu na duży stopień skomplikowania zagadnienia oraz jego potencjalnie rosnące znaczenie w przyszłych badaniach nad umysłem. Jak widać, sprowadzanie odpowiedzi na pytanie „Czy maszyny mogą myśleć” wyłącznie do optymistycznych ekstrapolacji na podstawie Prawa Moore’a, czy Prawa Kurzweil’a jest daleko idącym uproszczeniem.

Przedstawiona w tym rozdziale lista argumentów za i przeciwko możliwości powstania sztucznej inteligencji na ludzkim poziomie nie jest oczywiście wyczerpująca. Jednak już to zestawienie pokazuje, że odpowiedź na pytanie zadane w tytule nie jest jednoznaczna.

III. CZY MASZYNY POWINNY MYŚLEĆ?

„Prawdziwa głupota za każdym razem pokona sztuczną inteligencję.”

Terry Pratchett

Niektórzy wierzą, że kiedyś to sztuczna inteligencja pokona ludzką głupotę. „Gdy sterowane obliczeniowo roboty osiągną już nasz poziom inteligencji, zdaniem ekspertów nie będziemy musieli długo czekać, aż wysuną się błyskawicznie do przodu, daleko poza nasze skromne możliwości. Dopiero wtedy, twierdzą zwolennicy tego stanowiska, powstanie władza obdarzona dostateczną inteligencją, mądrością i zrozumieniem, by rozwiązać światowe problemy, które stworzyła ludzkość.”⁴² Z pozoru jest to bardzo atrakcyjne rozwiązanie, skoro SSI byłaby w stanie rozprawić się ze wszystkimi problemami trapiącymi człowieka, to stworzenie superinteligentnej maszyny byłoby ostatnim wynalazkiem potrzebnym ludzkości. Krok dalej idzie w swoich rozważaniach I.J.Good (1965), który twierdzi, że ultrainteligentna maszyna jest nam nie tylko potrzebna do rozwiązywania problemów, ale że to od jej skonstruowania zależy przetrwanie naszej rasy.

Jednak co, jeśli najlepszą receptą na ludzkie bolączki jest pozbycie się ludzkości? Czemu genialne maszyny miałyby nas słuchać, jeżeli nasz intelekt tak bardzo ustępowałby ich możliwościom? No i w końcu, czemu SSI miałyby być nam życzliwa?

Wracając do I.J.Gooda:

„Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is

⁴² Sir Roger Penrose *Cienie umysłu*, przeł. Piotr Amsterdamski, Poznań : Zysk i Sk-a Wydawnictwo, 2000,

sometimes worthwhile to take science fiction seriously.” („Tak więc pierwsza ultrainteligentna maszyna jest ostatnim wynalazkiem, jakiego człowiek będzie kiedykolwiek potrzebował, pod warunkiem, że maszyna będzie na tyle posłuszna, żeby powiedzieć nam jak utrzymać ją pod kontrolą. To zastanawiające, że ten problem jest tak rzadko poruszany poza fantastyką naukową. Czasem warto brać science fiction na poważnie.”)⁴³

Faktycznie, SF dostarcza wielu przykładów katastroficznych wizji związanych z rozwojem sztucznej inteligencji i nowoczesnych technologii w ogóle. *Matrix* braci Wachowskich przedstawia scenariusz supremacji maszyn, *Niezwyciężony* Lema opisuje wyprawę kosmonautów na planetę, na której znajdują się ślady cywilizacji zniszczonej przez nanotechnologiczną (inteligentną) chmurę. Nie brakuje też opisów zgubnych skutków stosowania inżynierii genetycznej, niektórych zaskakująco wczesnych jak *Dzień Tryfidów*(1951) Johna Wyndhama. Wizje fantastyki naukowej bywają mniej lub bardziej uzasadnione naukowo, ale z pewnością to właśnie temu gatunkowi zawdzięczamy zakorzenienie się tematu zagrożeń związanych z najnowszymi odkryciami naukowymi w zbiorowej wyobraźni. Można powiedzieć, że postulat I.J.Good'a został spełniony.

⁴³ I.J Good, *The First Ultraintelligent Machine*, Advances in Computers, t.6, 1965

III.1. CZŁOWIEK 2.0?

„Ubyliśmy zwierzętom.

Kto ubędzie nam.

Przez jakie podobieństwo.

Czego z czym porównanie.”

Wisława Szymborska, Notatka

III.1.1. CZEMU PRZYSZŁOŚĆ NAS NIE POTRZEBUJE

Sztuczna inteligencja (pod postacią robotów), nanotechnologia i inżynieria genetyczna występują jako potencjalne zagrożenia już nie tylko w powieściach i filmach SF. Bill Joy, współzałożyciel Sun Microsystems, wymienia triadę GNR (Genetics, Nanotechnology, Robotics) jako współczesny odpowiednik dwudziestowiecznych technologii nuklearnych, chemicznych i biologicznych, które doprowadziły do stworzenia broni masowego rażenia o niespotykanej wcześniej sile. Zdaniem Joy'a zagrożenia związane z technologiami dwudziestego pierwszego wieku są daleko gorsze od tych, z którymi ludzkość zmagająca się w przeszłości.

“I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals.”(„Myślę, że nie będzie wyolbrzymieniem stwierdzenie, że jesteśmy na progu czasu dalszego udoskonalania się ekstremalnego zła, zła, którego możliwości sięgają o wiele dalej, niż to, co zostawiły w spadku państwom narodowym bronie masowego rażenia, do zaskakującego i przerażającego wzmocnienia pozycji jednostek o skrajnych poglądach.”)⁴⁴

⁴⁴ Bill Joy, *Why The Future Doesn't Need Us*. “Wired” nr 8.04, 2000, [tłumaczenie własne]

Powodem do zaniepokojenia, zdaniem Bill'a Joy'a jest duża dostępność technologii GNR, mała kontrola nad ich rozwojem (ze względu na to, że badania prowadzone są w dużej mierze w komercyjnych centrach badawczych) oraz potencjalne zagrożenie samoreplikacją (robotów, nanobotów, efektów modyfikacji genetycznych). Do najczarniejszych scenariuszy należą m.in.: całkowite wyginięcie rodzaju ludzkiego na skutek ewolucyjnej wyższości robotów i pochłonięcie biosfery przez nanotechnologiczną „szarą maź” („grey goo”), w wyniku niekontrolowanego rozmnażania się nanobotów. Do tak katastrofalnych skutków miałyby doprowadzić zaawansowana technologia w niepowołanych rękach lub po prostu błąd, którego konsekwencje miałyby niespotykaną dotąd skalę.



Ilustracja III - 1: Napis na okładce: "Why the Future Doesn't Need Us By Bill Joy" („Czemu przyszłość nas nie potrzebuje”, Bill Joy) Wycinek: “human (hyŭ-mən) adj. 1. of, belonging to, or typical of the extinct species *Homo sapiens* <the human race> 2. what consisted of or was produced by *Homo sapiens* <human society> n. an extinct biped, *Homo sapiens*, characterized by carbon-based anatomy; also HUMAN BEING. Obs: hu•man•ness”, wycinek z okładki naśladuje hipotetyczne hasło z encyklopedii napisane o człowieku, już po jego wyginięciu. Źródło : <http://www.wired.com/wired/archive/8.04/>

Joy twierdzi, że powodem, dla którego ludzie nie przywiązują do tego typu zagrożeń dostatecznie dużej wagi, jest imperatyw postępu, ciekawość naukowców i zapotrzebowanie rynku na coraz szybsze i lepsze rozwiązania mające (w założeniu) ułatwiać ludzkie życie. Refleksja przychodzi tym trudniej, że mogłaby (jak chciałby Joy) doprowadzić do wniosku, że najrozsądniej byłoby nie rozwijać niektórych potencjalnie intratnych i interesujących z naukowego punktu widzenia projektów, ze względu na zbyt wysokie potencjalne zagrożenie jakie mogą nieść ze sobą one same lub ich skutki uboczne.

“Failing to understand the consequences of our inventions while we are in the rapture of discovery and innovation seems to be a common fault of scientists and technologists; we have been driven by the overreaching desire to know [...], not stopping to notice that the progress to newer and more powerful technologies can take a life of its own”(„Nieumiejętność dostrzeżenia konsekwencji naszych wynalazków kiedy jesteśmy w transie odkryć i innowacji wydaje się być powszechną wadą naukowców i technologów; jesteśmy powodowani sięgającym zbyt daleko pragnieniem, żeby wiedzieć [...], nie zatrzymując się, żeby zauważyć, że rozwój coraz to nowych i lepszych technologii może zacząć żyć własnym życiem”)

Nie mamy dzisiaj jednoznacznej odpowiedzi na pytanie czy maszyny mogą myśleć. Jest za to jeden sposób, żeby uzyskać pewność, że nigdy nie będą – wystarczy porzucić wszelkie próby stworzenia sztucznej inteligencji na ludzkim poziomie. Bill Joy jest zwolennikiem takiego właśnie ekstremalnego stanowiska – jeżeli istnieje ryzyko, że jakaś technologia przyczyni się do zagłady ludzkości, to należy ją porzucić, a być może nawet systemowo zakazać jej rozwijania. Oczywiście nie brakuje krytyków takiego rozwiązania. Sir Martin Rees⁴⁵, który częściowo podziela katastroficzne wizje Joy’a (szacuje, że szanse ludzkości na przetrwanie tego stulecia to 50:50⁴⁶) uważa takie radykalne rozwiązania za naiwne i niewykonalne. Rees proponuje zamiast tego promowanie odpowiedzialnej postawy u naukowców (a w szczególności akademików i niezależnych przedsiębiorców), którzy mieliby robić wszystko, żeby jak najlepiej informować o zagrożeniach związanych z rozwojem nowych technologii, wzorem fizyków zaangażowanych w projekt Manhattan ⁴⁷.

⁴⁵ Utytułowany brytyjski kosmolog i astrofizyk, autor książki *Our Final Century*

⁴⁶ Wystąpienie Martina Rees’a w ramach konferencji TEDGlobal2005.[online].[dostęp: 20.10.2013].Dostępny: http://www.ted.com/talks/martin_rees_asks_is_this_our_final_century.html

⁴⁷ Wiarę w dobrą wolę i altruistyczny wysiłek dużej grupy ludzi (w tym przypadku naukowców) też można uznać za naiwność. Szczególnie, że działania mające na celu uświadamianie społeczeństwa w kwestii potencjalnych skutków związanych z rozwijaniem technologii mogą obrócić się przeciwko samym uczonym, skutkując np. odebraniem im finansowania. Mimo to Martin Rees pokłada wiarę w inicjatywach takich jak Pugwash. Jak się okazuje, niebezpiecznym, ponieważ tematyka nanotechnologii i inżynierii genetycznej była już dyskutowana w ramach konferencji organizowanych przez Pugwash. Być może w niedalekiej przyszłości zagrożenia związane ze sztuczną inteligencją również staną się przedmiotem podobnego zainteresowania.

III.1.2. PASSING THE TORCH⁴⁸

Nie wszyscy naukowcy podzielają pesymizm Bill'a Joy'a. Hans Moravec, robotyk i futurysta pracujący w Instytucie Robotyki Carnegie Mellon University, zamiast mnożyć czarne scenariusze, skupia się na zaletach przyszłości zdominowanej przez roboty.

“Properly educated, the resulting robots will become quite formidable. In fact, I am sure they will outperform us in any conceivable area of endeavor, intellectual or physical. Inevitably, such a development will lead to a fundamental restructuring of our society. [...] it is likely that our descendants will cease to work in the sense that we do now. They will probably occupy their days with a variety of social, recreational and artistic pursuits, not unlike today's comfortable retirees or the wealthy leisure classes.” („Odpowiednio wyedukowane, roboty staną się całkiem potężne. Właściwie, jestem pewien, że przebiją nas w każdym możliwym obszarze działalności, intelektualnej czy fizycznej. Taki postęp będzie nieodzownie prowadził do fundamentalnej restrukturyzacji naszego społeczeństwa. [...] jest prawdopodobne, że nasi potomkowie przestaną pracować w takim sensie jak my dzisiaj. Prawdopodobnie będą wypełniali swój czas różnymi przedsięwzięciami towarzyskimi, rekreacyjnymi i artystycznymi, tak jak dzisiejsi żyjący w komforcie emeryci lub zamożne klasy niepracujące⁴⁹.”)⁵⁰

Moravec twierdzi, że jeżeli jego szacunki okażą się słuszne, to w 2050 roku większość faktów znanych nauce będzie zasługą odkryć robotów – naukowców, naszego „sztucznego potomstwa”⁵¹. Moravec używa tutaj terminu „potomstwo” nie bez powodu. W jego wizji superinteligentne roboty są naszymi dziećmi. Stworzonymi na nasze podobieństwo, a więc godnymi przejęcia schedy po ludzkości. Wizja Moraveca jest więc tylko pozornie utopijna, a właściwie jest utopią tylko jeżeli zgodzimy się na przyjęcie bardzo kontrowersyjnego założenia, że inteligentne roboty byłyby faktycznie przedłużeniem ludzkiej rasy, w ulepszonej formie.

⁴⁸ Ang. przekazanie pochodni. Hans Moravec stosuje to określenie w kontekście przekazania przez ludzi „pochodni ewolucji” komputerom

⁴⁹ ang. leisure class – klasa niepracująca/próżniacza

⁵⁰ Hans Moravec, *Rise of the Robots*, Scientific American, 1999

⁵¹ ang. artificial progeny

“Like biological children of previous generations, they [intelligent machines] will embody humanity’s best chances for long-term survival. It behooves us to give them every advantage and, when we have passed evolution’s torch, bow out.” („Tak, jak biologiczne dzieci poprzednich generacji, one [inteligentne maszyny] będą uosabiały największe szanse ludzkości na długoterminowe przetrwanie. Naszą odpowiedzialnością jest zapewnić im wszelkie korzyści i, kiedy przekazemy pochodnię ewolucji, wycofać się.”⁵²

Jak widać, Moravec prognozuje ludzkości rychłe przejście na emeryturę⁵³, która może się okazać całkiem komfortowa, jeżeli tylko zadbamy o to odpowiednio wcześniej. W praktyce oznacza to, że musimy „wychować” nasze robotyczne dzieci tak, żeby zechciały się nami w przyszłości zaopiekować.

Żeby potraktować wizję przyszłości nakreśloną przez Moraveca jako korzystną, a może nawet zbawienną dla ludzkości trzeba zgodzić się z założeniem, że roboty byłyby inteligentniejsze i sprawniejsze od nas. Wydaje się też, że musiałyby być nośnikami humanistycznych dążeń i wartości, przekazać dalej cały dorobek kulturowy człowieka. Moravec znajduje na to sposób – jego zdaniem maszyny będą zdolne do stworzenia modelu całej naszej cywilizacji, z dokładnością co do najdrobniejszego detalu, aż do poziomu atomowego⁵⁴. Udziela też prostej odpowiedzi na pytanie dlaczego maszyny miałyby coś takiego zrobić – otóż, jako ich twórcy będziemy dla nich interesujący... z historycznego punktu widzenia. Moravec nie przewiduje przetrwania ludzi w formie innej niż symulacja, nie traci jednak entuzjazmu (*“the takeover will be swift and painless”*⁵⁵). Co ciekawe, nie wyklucza, że być może już teraz żyjemy w symulacji, a roboty dawno już przeżyły oryginalną ludzkość (w końcu jeżeli maszyny miałyby stworzyć wiele symulacji naszej cywilizacji, to jest statystycznie bardziej prawdopodobne, że żyjemy właśnie w jednej z nich).

⁵² Hans Moravec, *Robots, Re-Evolving Mind at 10⁷ Times Nature’s Speed*, Cerebrum [online]. 2001, nr 2 [dostęp: 20.10.2013]. Dostępny: <http://www.dana.org/news/cerebrum/detail.aspx?id=3010>

⁵³ Nie tylko w przenośni – Moravec twierdzi, że w pełni zautomatyzowany przemysł zapewni ludzkości takie bogactwo, że nowonarodzeni będą mogli od razu przechodzić na emeryturę.

⁵⁴ *“They [machines] will, in fact, be able to re-create a model of our entire civilization, with everything and everyone in it, down to the atomic level, simulating our atoms with machinery that’s vastly subatomic”*

Hans Moravec: Superhumanism. Rozm. Przepr. Charles Platt. Wired, nr 3.10, 1995

⁵⁵ *“przejęcie będzie szybkie i bezbolesne”*

Tamże

Jak się okazuje, granica pomiędzy prognozą katastroficzną a utopijną może być bardzo cienka. Śmiałe wizje Moraveca znalazły się na ostrzu krytyki, między innymi autorstwa wspomnianego już w poprzednim rozdziale Roger'a Penrose'a.

III.2. PROBLEMY PRAKTYCZNE

„Wyśledzić moment historyczny, w którym liczydło dosięgło Rozumu, jest równie trudno jak ów, co małpę przemienił w człowieka.”

Stanisław Lem, „Golem XIV”

III.2.1. TRZY PRAWA ROBOTYKI

Zarówno Joy, Vinge, jak i Moravec prezentują skrajne wizje przyszłości. W ich kontrowersyjnych poglądach można jednak odnaleźć, w wyolbrzymionej formie, postulaty, które pojawiają się również w teoriach pozbawionych takiego rozmachu. Jednym z takich postulatów jest projektowanie inteligentnych robotów z myślą o ich przyszłej interakcji z ludźmi. Jak zaprojektować maszynę w taki sposób, żeby mieć pewność, że jej zachowanie nie wyrwie się spod kontroli, że niezależnie od procesu samodoskonalenia się maszyny, pewne podstawowe założenia pozostaną niezmienione? Po raz kolejny z pomocą przychodzą pisarze science fiction – w tym przypadku jeden z klasyków gatunku, Isaac Asimov, który w swoich książkach opisał roboty wyposażone w zestaw podstawowych, niepodważalnych praw regulujących ich działanie⁵⁶:

„Trzy Prawa Robotyki:

1. Robot nie może wyrządzić krzywdy istocie ludzkiej lub przez bezczynność pozwolić aby stała się jej krzywda.

⁵⁶ Prawo zerowe przytaczam po prawach 1-3, ze względu na jego późniejsze pojawienie się w chronologii książek Asimova.

2. Robot musi słuchać rozkazów wydawanych mu przez istoty ludzkie, z wyjątkiem sytuacji, w których takie rozkazy byłyby sprzeczne z Pierwszym Prawem.

3. Robot musi chronić swoje istnienie dopóty, dopóki taka ochrona nie jest sprzeczna z Pierwszym lub Drugim Prawem.”⁵⁷

„0. Robot nie może zaszkodzić ludzkości czy też przez niepodjęcie działania pozwolić, by ludzkość poniosła szkodę.”⁵⁸

Oryginalne trzy prawa robotyki Asimova powstały jeszcze w latach czterdziestych, a więc na długo przed tym, zanim problematyka etyki robotów mogła być przedmiotem realnego zainteresowania. Są zgodne z intuicją – mogłoby się wydawać, że zaprogramowanie takich ograniczeń faktycznie zabezpieczyłoby ludzi przed niepożądanym zachowaniem sztucznej inteligencji. W rzeczywistości jednak, tak sformułowane zasady mogłyby, potencjalnie, prowadzić do wielu komplikacji. Przykładowo: robot nie może pozwolić, żeby człowiekowi stała się krzywda poprzez wstrzymanie się od działania. Nie jest jednak określone jak dalece robot musiałby rozważać potencjalne konsekwencje wszystkich swoich możliwych zaniechań. Niewykluczone, że wnioskowanie robota o podjęciu, lub niepodjęciu akcji opierałoby się na rozważaniu wszystkich możliwych scenariuszy rozwoju sytuacji, a więc wszystkich tych rzeczy, które mogą się w danej sytuacji wydarzyć, jak również tych, które nie będą mieć miejsca. Realizacja takiego założenia zależałaby w takim przypadku od możliwości rozwiązania problemu ramy odniesienia (II.2.2.1.). Z drugiej strony, pominięcie w pierwszym prawie fragmentu „poprzez wstrzymanie się od działania” mogłoby w oczywisty sposób prowadzić do tragicznych konsekwencji.

Przede wszystkim jednak, prawa robotyki Asimova wydają się być zbiorem reguł zbyt nieskomplikowanym, żeby być odpowiednimi dla sterowania zaawansowanymi systemami AI. Stephen Omohundro w artykule „The Basic

⁵⁷ Isaac Asimov, *Lustrzane odbicie* [w:] *Świat robotów*, przeł. Edward Szmigiel, Varia APD, 1993

⁵⁸ Isaac Asimov, *Roboty i Imperium*, przeł. Paulina Braiter, Wyd. 2 popr., Poznań: Dom Wydawniczy Rebis, 2003

AI Drives” zwraca uwagę na stopień trudności stojącego przed ludźmi zadania stworzenia „bezpiecznej” sztucznej inteligencji.

“Surely no harm could come from building a chess-playing robot, could it? [...] Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else’s safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of any goal driven systems.” (“Z pewnością nic złego nie może wyniknąć ze zbudowania robota grającego w szachy, prawda? [...] Jeżeli nie zachowamy specjalnych środków ostrożności, będzie opierał się próbom wyłączenia go, będzie próbował włamać się do innych maszyn i tworzyć kopie siebie, oraz będzie starał się zdobywać zasoby bez względu na czyjekolwiek bezpieczeństwo. Te potencjalnie szkodliwe zachowania nie pojawią się dlatego, że zostały zaprogramowane na początku, ale dlatego, że taka jest nieodłączna natura każdego systemu zorientowanego na cel.”)

Omohundro rozważa cele, które mogą, jego zdaniem, powodować zachowaniem inteligentnych maszyn. Wszystkie one odnoszą się, w ten lub inny sposób, do funkcji użyteczności⁵⁹, która ma definiować cele robota

Wg. Omohundro AI będą chciały:

- 1) pracować nad samodoskonaleniem
- 2) być racjonalne
(żeby móc wnioskować o niepewnej przyszłości)
- 3) zachować swoją funkcję użyteczności
(funkcja użyteczności ma stanowić, w pewnym sensie, odpowiednik ludzkiego zbioru celów i wartości, dlatego też ich drastyczna zmiana

⁵⁹ Funkcja użyteczności każdemu stanowi maszyny/programu przypisuje wartość użyteczności tego stanu. W przypadku, kiedy agent operuje w świecie rzeczywistym (albo środowisku, w którym występują zmienne losowe oraz częściowa obserwowalność) musi posługiwać się oczekiwaną użytecznością (ang. expected utility) danego stanu. Zadaniem działającego racjonalnie agenta jest maksymalizacja własnej funkcji użyteczności.

(np. na skutek wrogiej ingerencji) powinna być dla maszyny druzgocącą perspektywą⁶⁰)

4) zapobiegać fałszywej użyteczności

(maszyny będą potrzebowały dobrej wewnętrznej reprezentacji swoich celów, żeby uniknąć działań na rzecz „fałszywej użyteczności”; przykładowo, program, który ma za zadanie grać w szachy potrzebuje takiej funkcji użyteczności, która wykluczy zaistnienie sytuacji, w której wcale nie będzie on grał, a jedynie zwiększał w nieskończoność licznik wygranych partii)

5) chronić siebie

6) zdobywać zasoby i efektywnie je wykorzystywać

(większa ilość zasobów zwiększa oczekiwaną użyteczność)⁶¹

Każde z przedstawianych przez Omohundro potencjalnych dążeń robotów zwiększa trudność nałożenia na maszyny ograniczeń mających zapewnić nam bezpieczeństwo (podobnie jak u ludzi – im więcej czynników, które nami kierują, tym trudniej przewidzieć nasze zachowanie). Jeżeli roboty przyszłości miałyby być naprawdę inteligentne i zdolne do samomodyfikacji, to każda próba zastosowania zewnętrznych restrykcji np. w stosunku do możliwości samoulepszenia, musiałaby spotkać się z próbami jej obejścia, w celu zwiększenia wartości spodziewanej użyteczności. Co więcej, sama maszyna może mieć nieomal ludzkie problemy z rozwiązywaniem konfliktów między własnymi celami.

Omohundro postuluje też stworzenie stosownego kontekstu społecznego dla robotów:

“I believe that we should begin designing a “universal constitution” that identifies the most essential rights we desire for individuals and creates social mechanisms for ensuring them in the presence of intelligent entities of widely varying structures”. („Wierzę, że powinniśmy zacząć projektować “uniwersalną konstytucję”, która identyfikowałaby najbardziej podstawowe

⁶⁰ Można powiedzieć, że wartość funkcji użyteczności dla danego stanu odpowiada stopniowi „zadowolenia” maszyny

⁶¹ Tytuły punktów odpowiadają tytułom podrozdziałów artykułu Stephena Omohundro („*The Basic AI Drives*”)

prawa, których pragniemy dla jednostek i definiowała mechanizmy społeczne, mające zapewnić niezbywalność tych praw w obecności inteligentnych jednostek o bardzo różnej strukturze”)

Jak widać, zadanie stworzenia sensownych ograniczeń funkcjonowania robotów może się okazać co najmniej tak skomplikowane jak tworzenie ludzkiego prawodawstwa. Nie wystarczą tu niestety 3, ani nawet 4 podstawowe prawa.

III.2.2. KTO OSĄDZI SZTUCZNĄ INTELIGENCJĘ?

Tworzenie „prawa dla robotów” może wydawać się, póki co, mało praktycznym zajęciem, jednak wbrew pozorom, sztuczna inteligencja, nawet w obecnej formie, jest już zdolna wywoływać prawne kontrowersje. Najlepiej świadczy o tym istnienie takich periodyków jak *Artificial Intelligence and Law* („Sztuczna Inteligencja i Prawo”), czy też odbywające się regularnie konferencje na ten temat (pierwsza konferencja *AI and Law* odbyła się w Bostonie już w 1987 roku).

Przykładem tego, jak sztuczna inteligencja niepostrzeżenie staje się częścią rzeczywistości (nie tylko prawnej) są samochody sterowane przez systemy AI. Słynne auto Google bynajmniej nie jest pierwszym tego typu przedsięwzięciem. Przykładowo, samosterujący się samochód NavLab V⁶² przemierzał Stany Zjednoczone już w 1995 roku, radząc sobie bez ludzkiej pomocy przez 98% czasu. Zarówno NavLab 5 (i jego obecna wersja, NavLab 11), jak i auto Google nie wyróżniają się zbytnio swoim wyglądem zewnętrznym.

⁶² Zbudowany w Instytucie Robotyki Carnegie Mellon University NavLab V w 1995 roku pokonał niecałe 5 tysięcy kilometrów (3 tysiące mil) w ramach trasy “No Hands Across America”.



Ilustracja III - 2 Auto Google, zaadaptowany Lexus RX450 h, na dachu widoczny radar LIDAR
źródło:<http://tctechcrunch2011.files.wordpress.com/2012/08/self-driving-car-google-logo.jpg?w=400>

Głównym elementem odróżniającym Google Car od innych Lexusów RX450h jest umiejscowiony na dachu trójwymiarowy skaner otoczenia – LIDAR. Być może już niedługo będzie można zupełnie nieświadomie minąć je po drodze do pracy. W 2012 roku Eric Schmidt, były prezes zarządu firmy Google, przewidywał, że samosterujące się samochody staną się dominującym środkiem transportu jeszcze za naszego życia. Z kolei w sierpniu bieżącego roku Nissan ogłosił, że będzie produkował (w przystępnej cenie) samosterujące się auta, już w 2020 roku.⁶³

Samosterujące się samochody nie są jednak wyłącznie problemem technologicznym (zresztą Google testuje już swój wynalazek w warunkach ruchu ulicznego, a jeśli wierzyć plotkom, także na torach wyścigowych), ale również prawnym. Obecnie, na skutek działań Google, samosterujące się samochody mogą legalnie przemierzać drogi Kalifornii i Nevady (aczkolwiek istnieje szereg warunków, które muszą spełnić). Biorąc pod

⁶³ Cytując słowa prezesa firmy, Carlosa Ghosn'a:

"In 2007 I pledged that—by 2010—Nissan would mass market a zero-emission vehicle. Today, the Nissan LEAF is the best-selling electric vehicle in history. Now I am committing to be ready to introduce a new ground-breaking technology, Autonomous Drive, by 2020, and we are on track to realize it." ("W 2007 roku ogłosiłem, że – do 2010 roku – Nissan będzie prowadził masową produkcję auta o zerowej emisji spalin. Dziś, Nissan LEAF jest najlepiej sprzedającym się samochodem elektrycznym w historii. Teraz, zobowiązuję się być gotowy do wprowadzenia nowej, przełomowej technologii, Autonomous Drive, do 2020 roku, i jesteśmy na dobrej drodze, żeby to zobowiązanie zrealizować."), [tłumaczenie własne], cytat przytaczam za CNBC:

Phil LeBeau, *"Nissan to build self-driven cars"*. [online]. [dostęp: 20.10.2013]. Dostępny: <http://www.cnbc.com/id/100991243>

uwagę dotychczasową skuteczność lobbingu Google na rzecz swojego wynalazku, w niedługim czasie można spodziewać się legalizacji tego typu wehikułów w innych stanach Stanów Zjednoczonych.

Sama kwestia legalizacji nie jest jednak jedynym potencjalnym problemem. Cytując raport CAR (Center of Automotive Research) i firmy KPMG LLP:

"If the driver, by design, is no longer in control, what happens if the vehicle crashes? The "driver" could well be an innocent bystander or might at least bear lesser liability than drivers do today. A legal framework will be necessary to deal with the potentially complex liability issues that may come with self-driving cars."

(„Jeżeli kierowca, zgodnie z założeniem, jest pozbawiony kontroli, co się stanie jeżeli samochód się rozbije? „Kierowca” mógłby równie dobrze być niewinnym świadkiem lub mógłby przynajmniej ponosić mniejszą odpowiedzialność niż obecnie. Potrzebny będzie kontekst prawny, żeby radzić sobie z potencjalnie skomplikowanymi problemami z odpowiedzialnością, które mogą wyniknąć przy samosterujących się autach.”)⁶⁴

Podobne problemy mogą dotyczyć również innych dziedzin sztucznej inteligencji. Przykładem mogą być systemy ekspertowe – kto będzie odpowiedzialny za złą diagnozę opartą na zaleceniach takiego systemu, lekarz, system, programista?

Jeszcze ciekawsze problemy mogą pojawić się kiedy samosterujące się samochody będą jeździć lepiej niż prawdziwi kierowcy, a systemy ekspertowe zaczną osiągać trafniejsze wyniki niż sami eksperci. Być może wtedy ludzie będą musieli odpowiadać za *nie* skorzystanie z pomocy lub rady maszyny. Niewykluczone, zresztą, że będzie ich wtedy sądził nie człowiek, a komputer – taką ewentualność (poza sądami apelacyjnymi) dopuszcza Joseph S. Fulda w artykule *„Implications of a logical paradox for*

⁶⁴ Raport: *“Self-driving cars: The next revolution”*. [online]. [dostęp 27.10.2013]. Dostępny: <http://www.kpmg.com/Ca/en/IssuesAndInsights/ArticlesPublications/Documents/self-driving-cars-next-revolution.pdf>

computer-dispensed justice reconsidered: some key differences between minds and machines". Fulda uznaje, że maszyny (o ile tylko osiągną poziom skuteczności ludzkich sędziów, lub przynajmniej się do niego zbliżą) mogłyby, a nawet powinny oceniać czy ktoś jest winny, czy nie. Trzeba jednak zaznaczyć, że Fulda nie dopuszcza wydawania przez maszynę wyroków⁶⁵.

III.2.3. RACE AGAINST THE MACHINE⁶⁶

Prawdopodobnie najbardziej rozpowszechnione obawy związane z rozwojem sztucznej inteligencji dotyczą potencjalnego zmniejszenia się liczby miejsc pracy. Nawet gdyby wprowadzenie na szeroką skalę robotycznych pracowników miało w długim okresie czasu zaowocować utopią, w której ludzie przechodziliby na emeryturę już w dniu urodzenia, to nadal w krótkim okresie czasu mogłoby to doprowadzić do masowego bezrobocia, zwłaszcza wśród ludzi gorzej wykształconych. Wprawdzie jak na razie obserwujemy raczej wzrost liczby miejsc pracy w wyniku postępującej komputeryzacji i mechanicyzacji, ale nie wiadomo jak sytuacja będzie się rozwijać wraz z dalszym rozwojem technologii. Przykładowo, skuteczność programów do tłumaczenia tekstów technicznych jest już tak duża, że niektóre firmy planują w niedługiej przyszłości zastąpienie nimi ludzkich tłumaczy. Z kolei potentat internetowej sprzedaży, portal Amazon,

⁶⁵ Fulda argumentuje, że wszystko to, co czyni maszynę odpowiednią do opiniowania o winie, sprawia, że nie nadaje się ona do wydawania wyroków (m.in. ze względu na konieczność daleko idącej personalizacji wyroków); naszelną cechą, której ma potencjalnie brakować maszynie jest mądrość:

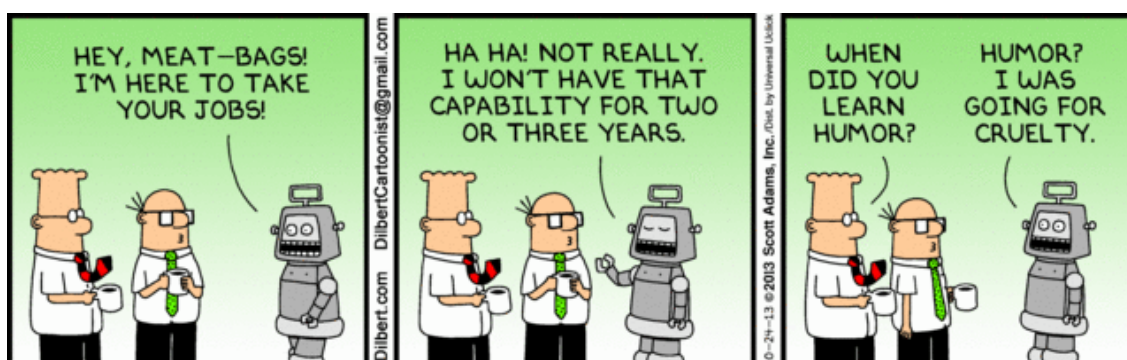
"However, intelligence [...] may be all that is needed to decide guilt or innocence, but it is hardly enough to decide what to do with a person once found guilty. This requires an entirely different quality, wisdom. I do not know how to teach a machine wisdom, nor how to program one to learn to be wise—by which I mean that I am not aware of research aiming at this elusive goal; it is hard even to define this quality."(Jakkolwiek, inteligencja [...] może być wszystkim, co jest potrzebne do zadecydowania o winie lub niewinności, to jednak nie jest ona wystarczająca do zdecydowania co zrobić z osobą, która została uznana za winną. To wymaga zupełnie innej jakości, mądrości. Nie wiem jak nauczyć maszynę mądrości, ani jak zaprogramować taką, która nauczy się być mądrą – przez co rozumiem, że nie mam wiedzy o badaniach, które miałyby prowadzić do osiągnięcia tego ulotnego celu; ta jakość jest trudna nawet do zdefiniowania") Joseph S. Fulda, „*Implications of a logical paradox for computer-dispensed justice reconsidered: some key differences between minds and machines*”, *Artificial Intelligence and Law*, 2012, t. 20, nr 3, s 321-333

Joseph Fulda porusza tutaj rzadko rozpatrywaną kwestię – czy da się zaprogramować sztuczną mądrość?

⁶⁶ Tytuł książki Erika Brynjolfsson'a i Andrew McAfee, dosłownie wyścig przeciwko maszynie. Tytuł jest grą słów nawiązującą do nazwy zespołu Rage against the Machine.

opracował na tyle dobry algorytm książkowych rekomendacji (służący do sugerowania kupującemu pozycji, które mogą mu się potencjalnie spodobać), że przewyższył on w skuteczności ludzkich pracowników, w konsekwencji czego ich etaty zostały zredukowane⁶⁷.

Utrata miejsc pracy może wiązać się z większą efektywnością działania robotów/programów sztucznej inteligencji. Może się jednak także wiązać z takim wzrostem skomplikowania technologii, że będzie ona wykraczać daleko poza możliwości ludzkiego zrozumienia. Już w tej chwili można obserwować podobne zjawisko; dla wielu ludzi, jak ujął to Arthur C. Clarke, „Każda wystarczająco zaawansowana technologia jest nieodróżnialna od magii”.



Ilustracja III - 3 Motyw utraty pracy na skutek rozwoju sztucznej inteligencji jest rozpowszechniony w kulturze popularnej.

Źródło: <http://www.dilbert.com/2013-10-24/>

Kolejnym powodem do strachu przed sztuczną inteligencją może być utrata poczucia wyjątkowości – na zasadzie podobnej do przewrotu kopernikańskiego. Tak jak Ziemia przestała być centrum wszechświata, tak, jeżeli uda się stworzyć sztuczną inteligencję na ludzkim poziomie, ludzie przestaną być unikalni. Można jednak argumentować, że skoro zmiana układu odniesienia na heliocentryczny nie wpłynęła druzgocąco na kondycję człowieka, to również istnienie zaawansowanych programów sztucznej inteligencji zostanie, w swoim czasie, zaakceptowane i będzie traktowane jako coś zupełnie normalnego. Z drugiej strony, o ile perspektywa kosmiczna jest dla przeciętnego człowieka (przynajmniej na co dzień) dość obojętna, o tyle poczucie nieużyteczności w porównaniu

⁶⁷ Informację podają za: Sue Halpern, „*Are We Puppets in a Wired World?*”, *The New York Review of Books*, Listopad 2013

z maszynami i braku celu w życiu, które może być powodowane brakiem konieczności pracy, może w poważny sposób zaważyć na kondycji przyszłych społeczeństw.

Podsumowanie

Powyższy rozdział rozpoczyna się od katastroficznych wizji i hipotetycznych problemów związanych z rozwojem sztucznej inteligencji, a kończy na kłopotach bardziej przyziemnych i występujących już dzisiaj. Zarówno pierwsze, jak i drugie mogą służyć jako poparcie przytaczanego przez Turinga argumentu „głów w piasku” (II.2.1.2.). Oczywiście, część niepokoju związanego z rozwojem sztucznej inteligencji i nowych technologii w ogóle, jest związana z technofobią, która z kolei jest potęgowana przez kulturę popularną, która chętnie żywi się ludzkim strachem i neurozami.

Wydaje się, że jakkolwiek części przewidywanych zagrożeń związanych z rozwojem sztucznej inteligencji nie da się zredukować do uprzedzeń wynikających ze strachu przed technologią, to radykalne propozycje porzucenia badań nie są tutaj najlepszym, ani możliwym do realizacji, rozwiązaniem. Argument „głów w piasku” można więc odrzucić ze względu na sam fakt niemożliwości jego aplikacji. Przyjęcie kompromisowego stanowiska Martina Rees’a, wymaga jednak dużej uwagi ze strony badaczy. Przychylając się częściowo do postulatów Moraveca o „wychowaniu” sobie posłusznych „dzieci umysłu”, tworząc sztuczne jednostki autonomiczne, trzeba zadbać, aby nie były one autonomiczne etycznie. ⁶⁸

⁶⁸ “We should also ask what might happen if a child somehow gained more control [...] – so that now that *child could praise himself*-and thus select which goals to elevate – or else *that child could censure himself*, and thus impose new constraints on himself. This would make him “ethically autonomous”[...]”

(„Powinniśmy też zapytać, co może się zdarzyć, jeżeli dziecko w jakiś sposób uzyska większą kontrolę – tak, że teraz *dziecko będzie mogło chwalić siebie* – a przez to wybierać które cele uwznioślić”)

“[I]f that child’s brain were able to change all of this previous values and goals, then there would remain no constraints at all on what kind of person emerges from this – it could even be a sociopath.”

(“Jeżeli mózg tego dziecka będzie zdolny do zmiany wszystkich swoich poprzednich celów i wartości, wtedy nie będzie żadnych ograniczeń odnośnie tego jaka osoba się z tego wyłoni – to może nawet być socjopata”) [tłumaczenia własne]

“Finally, it seems likely that a large-scale success in AI – the creation of human-level intelligence and beyond - would change the lives of a majority of humankind. The very nature of our work and play would be altered, as would our view of intelligence, consciousness, and the future destiny of the human race. AI systems at this level of capability could threaten human autonomy, freedom, and even survival. For these reasons, we cannot divorce AI research from its ethical consequences.” („Ostatecznie, wydaje się, że sukces SI na dużą skalę – stworzenie sztucznej inteligencji na ludzkim poziomie i lepszej – zmieni życie większej części ludzkości. Zmieni się natura naszej pracy i zabawy, tak jak nasze spojrzenie na inteligencję, świadomość i przyszłość ludzkiej rasy. Systemy AI na tym poziomie możliwości mogą zagrozić ludzkiej autonomii, wolności, a nawet przetrwaniu. Z tych powodów, nie możemy oddzielić badań nad sztuczną inteligencją od jej etycznych konsekwencji.”)⁶⁹

Minsky używa tutaj określenia “autonomii etycznej” w kontekście ludzkich dzieci, ale jego uwagi mogą mieć praktyczne zastosowanie dla dzieci elektronicznych.

Marvin Minsky, *“The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and The Future of the Human Mind”*, Nowy York: SIMON&SCHUSTER PAPERBACKS, 2006

⁶⁹ Stuart Russel, Peter Norvig, *„Artificial Intelligence: A Modern Approach”*, New Jersey: Pearson Education, 2010

IV. THE IMITATION GAME

"But how are we to tell when we have reached our destination? Turing's test is the only game in town"

*The Oxford Handbook of Philosophy of Cognitive Science.*⁷⁰

W jaki sposób sprawdzić, czy cel został osiągnięty? Przyjmując założenie, że sztuczna inteligencja na ludzkim poziomie może istnieć, w jaki sposób upewnić się, że jeszcze nie powstała? W końcu, jeśli zgodzić się z autorami najczarniejszych scenariuszy dotyczących rozwoju technologii, taka pewność jest jedynym gwarantem przetrwania ludzkiej rasy. Na jakie wyniki w jakim teście powinni z utęsknieniem czekać entuzjaści powstania SSI, transhumaności⁷¹?



Ilustracja IV - 1: Medal dla zwycięzcy głównej nagrody konkursu o nagrodę Loebnera, wykonany w całości ze złota (po drugiej stronie podobizna fundatora nagrody),

źródło: <http://www.loebner.net>

Okazuje się, że mimo upływu lat najbardziej znaną i właściwie jedyną liczącą się propozycją takiego testu pozostaje test Turinga, nazwany tak na cześć swojego autora, a zaproponowany po raz pierwszy w cytowanym już kilkakrotnie artykule „*Computing Machinery and Intelligence*”. Mimo powstania różnych wersji tego testu, jak również

jego zdecydowanej krytyki przez część środowiska naukowego, co roku programy AI walczą o zwycięstwo w konkursie o nagrodę Loebnera, gdzie ich zadaniem jest właśnie przejście testu Turinga.

⁷⁰ „Jak stwierdzić, czy osiągnęliśmy cel? Test Turinga jest jedynym, co mamy do wyboru.”, [tłumaczenie własne]

⁷¹ Transhumanizm jest ruchem opartym na założeniu, że dzięki nauce i technice ludzie będą w stanie przewyciężyć swoje ograniczenia (przykładowo wyeliminować proces starzenia się), a nawet wykroczyć poza swoją pierwotną fizyczną formę (np. „przepisać się „ do postaci cyfrowej)

Pomimo argumentów swoich przeciwników i ponad sześćdziesięciu lat kariery, TT⁷² nie traci oddziaływania na zbiorową wyobraźnię. Pozostaje pytanie czy słusznie i czy w najbliższym czasie jakimkolwiek programowi lub maszynie uda się spełnić warunki postawione przez Turinga.

IV.1. TEST NA CZŁOWIECZEŃSTWO

"May not machines carry out something which ought to be described as thinking but which is very different from what a man does? [...] This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection."

*Alan Turing*⁷³

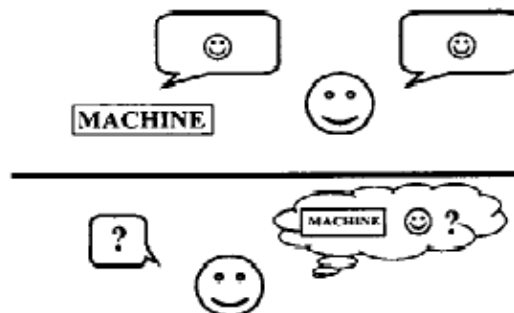
IV.1.1. DEFINICJA

Turing zaproponował by pytanie „Czy maszyny mogą myśleć” zastąpić pytaniem czy maszyny są w stanie zdać wymyślony przez niego test, który z kolei jest adaptacją gry towarzyskiej, *imitation game* (ang. gra w naśladownictwo). Uczestnikami są tutaj: dwoje ludzi (gracz 1 i sędzia) oraz komputer (gracz 2). Zadaniem każdego z graczy jest przekonanie sędziego o swoim człowieczeństwie w trakcie normalnej rozmowy (prowadzonej, dla wyrównania szans, w trybie tekstowym). Tak więc, maszyna zwycięży, jeżeli będzie w stanie udawać człowieka bardziej przekonująco niż sam człowiek, czyli gracz 2.

⁷² W dalszej części rozdziału będę stosować na określenie testu Turinga skrót „TT”

⁷³ „Czy maszyna nie mogłaby przeprowadzać czegoś, co powinno być nazwane myśleniem, ale w sposób różny od tego, w jaki robi to człowiek? [...] Ten sprzeciw jest bardzo silny, ale możemy przynajmniej powiedzieć, że jeżeli, pomimo wszystko, maszyna może zostać skonstruowana tak, by grać w grę w naśladownictwo w satysfakcjonujący sposób, to nie musimy się tym zarzutem przejmować.”

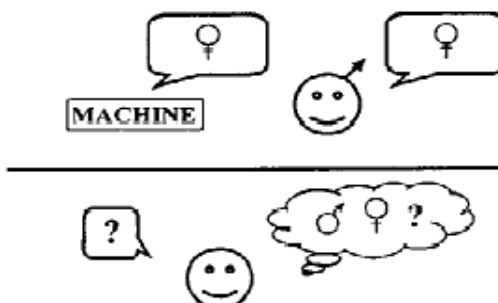
Alan Turing, *Computing Machinery and Intelligence*, [tłumaczenie własne]



Ilustracja IV - 2: Test Turinga.

Źródło: Ilustracja pochodzi z publikacji "Turing Test: 50 Years Later", Ayse Pignar Saygin, Ilyas Cicekli & Varol Akman

Co ciekawe, w oryginalnej wersji gry w naśladownictwo, gracze (gracz 1 będący w tym przypadku mężczyzną, gracz 2 - komputerem) mieli udawać nie maszynę, a kobietę. W takim przypadku, każdy z graczy byłby zmuszony do udawania. W ostatecznej wersji natomiast ludzki gracz jest w uprzywilejowanej pozycji – udawanie samego siebie nie wymaga bowiem wielkiego namysłu.



Ilustracja IV - 3: Pierwotny kształt imitation game. Mężczyzna i maszyna konkurują w udawaniu kobiety.

Źródło: Ilustracja pochodzi z publikacji "Turing Test: 50 Years Later", Ayse Pignar Saygin, Ilyas Cicekli & Varol Akman

Wydaje się, że takie sformułowanie testu jest dość niesprawiedliwe dla maszyny. Sam Turing zauważa zresztą, że w przypadku odwrócenia testu (kiedy człowiek i maszyna rywalizowałyby w udawaniu maszyny), ludzki zawodnik miałby przed sobą bardzo trudne zadanie - *"If the man were to try and pretend to be the machine he would clearly make a very poor showing"* („Jeżeli człowiek miałby spróbować

udawać maszynę, to oczywiście jego osiągnięcia byłyby bardzo słabe”)⁷⁴. Turing nie oczekuje jednak, że maszyna musiałaby wygrywać za każdym razem, aby udowodnić swoją inteligencję:

"I believe that in about fifty years' time it will be possible to programme computers, with a storage capacity of about 10^9 , to make them play the imitation game so well that an average interrogator will not have more than 70 per cent. chance of making the right identification after five minutes of questioning"(Wierzę, że za mniej więcej pięćdziesiąt lat będzie możliwe zaprogramowanie komputerów, o pojemności pamięci około 10^9 , do gry w naśladownictwo na tyle dobrze, że przeciętny rozmówca nie będzie miał więcej niż 70 procent szans na właściwą identyfikację po pięciu minutach przepytania)⁷⁵

Trzydzieści procent szans na wygraną może nie wydawać się dużą liczbą, ale jak dotychczas żadnej maszynie nie udało się spełnić takiego wymagania, mimo że dzisiejsze komputery osobiste dysponują już dyskami twardymi o pojemności liczonej w terabajtach.

IV.1.2. WERSJA WSPÓŁCZESNA

W 1990 roku kontrowersyjny amerykański wynalazca Hugh Loebner⁷⁶ ufundował nagrodę, w postaci złotego medalu (przedstawionego na ilustracji na początku rozdziału) oraz stu tysięcy dolarów, dla „pierwszego komputera, którego odpowiedzi będą nieodróżnialne od ludzkich” („[...] for the first computer whose responses were indistinguishable from a human's”). Srebrny medal, oraz dwadzieścia pięć tysięcy dolarów to nagroda przewidziana dla autora programu zdolnego oszukać połowę sędziów. Dodatkowo, corocznie przyznawana jest nagroda dla „najbardziej ludzkiego komputera” („most human computer”), w postaci brązowego medalu i nagrody pieniężnej w wysokości kilku tysięcy dolarów.

⁷⁴A.Turing,[tłumaczenie własne]

⁷⁵ *Tamże*

⁷⁶ Hugh Loebner jest ekscentrykiem i orędownikiem legalizacji prostytucji, zarzuca mu się, że ufundował nagrodę tylko ze względu na chęć zdobycia większego rozgłosu.

Formuła konkursu nie jest jednak stała. W momencie, w którym zostanie przyznany srebrny medal zmieniają się wymagania wobec uczestników – w następnym roku po przyznaniu srebrnego medalu uczestniczące w konkursie programy będą musiały przetwarzać multimodalne⁷⁷ informacje na wejściu, a więc nie tylko tekst, ale również programy, filmy, etc. Jeżeli w fazie multimodalnej któremuś z programów ponownie uda się sztuka oszukania połowy sędziów, to jego autorowi zostanie przyznany złoty medal, a konkurs nie będzie już więcej organizowany.⁷⁸

Jak widać, złoty medal w rzeczywistości nie przypadnie autorowi programu, który spełni warunki przedstawione przez Turinga, ale takiemu, który zda jego rozszerzoną wersję. Póki co jednak, żadna nagroda poza dorocznym brązowym medalem nie została przyznana, sędziowie oceniają więc inteligencję programów jedynie na podstawie rozmowy prowadzonej na ekranie komputera.

Co roku, w ramach finału konkursu, czterech⁷⁹ sędziów ocenia umiejętności czterech programów, które osiągnęły najwyższe wyniki podczas kwalifikacji. Często nie są oni, wspomnianymi przez Turinga, „przeciętnymi rozmówcami”, tylko ekspertami w dziedzinie sztucznej inteligencji czy psychologii.⁸⁰ Maszyny w trakcie finału muszą konkurować z prawdziwymi ludźmi. Podczas każdej z czterech finałowych rund jeden sędzia rozmawia jednocześnie z jednym człowiekiem i jednym komputerem. Jego zadaniem jest zdecydowanie który z jego rozmówców jest maszyną. Można więc powiedzieć, że w trakcie finału zachowana jest formuła testu zaproponowana przez Turinga.

Inaczej rzecz ma się w trakcie kwalifikacji, kiedy to sędziowie oceniają jak „ludzkie” są ich zdaniem odpowiedzi udzielone przez zgłoszone programy. W procesie kwalifikacji nie ma porównania odpowiedzi człowieka z odpowiedziami generowanymi przez maszynę. Poniżej przytaczam przykładowe pytania kwalifikacyjne (na rok 2011):

⁷⁷ „MultiModal”

⁷⁸ Informacje i cytaty pochodzą z oficjalnej strony konkursu: <http://www.loebner.net/Prize/loebner-prize.html>, dostęp: 30.10.2013

⁷⁹ Przeważnie, ale liczba ta nie jest stała.

⁸⁰ Przykładowo: jednym z sędziów w 2011 roku był Professor Noel Sharkey, Professor of Artificial Intelligence & Robotics z University of Sheffield

1. *My name is Ed. What is your name?*
2. *Which is larger, an ant or an anteater?*
3. *What month of the year is it?*
4. *What is my name?*
5. *Dave is older than Steve but Steve is older than Jane. Who is youngest, Steve or Jane?*
6. *What day will it be tomorrow?*
7. *What's your favorite food?*
8. *My friend Chris likes to play football. What sports do you like to play?*
9. *Are you a human or a computer?*
10. *What letter comes after T?*
11. *What is the name of my friend who likes to play football?*
12. *What is your name?*
13. *What would I do with a screwdriver?*
14. *How many letters are in the word 'banana'?*
15. *Have you watched a good film lately?*
16. *What year will it be next year?*
17. *What's your favorite fruit?*
18. *Do you prefer white or black coffee?*
19. *How old are you?*
20. *The football was kicked by Fred. Who kicked the football?⁸¹*

Powyższe pytania mogą wydawać się proste, jednak okazuje się, że stanowiły one pewne wyzwanie dla uczestników konkursu, przykładowo Rosette, program autorstwa Bruce'a Wilcox'a, który osiągnął najwyższy wynik w kwalifikacjach (a później był zdobywcą brązowego medalu), na pytanie „What letter comes after T?” udzielił odpowiedzi „I don't know.”, natomiast na pytanie „What year will it be next year?” udzielił dość absurdalnej odpowiedzi „Dave.”.

O ile wyniki kwalifikacji wyglądają nieraz dość obiecująco, to entuzjaści sztucznej inteligencji mogą być zawiedzeni wynikami finałów. Przykładem może być tutaj fragment rozmowy z jednym z programów zakwalifikowanych do finału przytaczany przez sędziego, profesora Noel'a Sharkey'a w artykule dla The Guardian (wypowiedzi sędziego są oznaczone pogrubieniem czcionki):

“Could you kill someone if you stabbed them with a towel?”

⁸¹ „1. Mam na imię Ed. Jak masz na imię? 2. Co jest większe, mrówka czy mrówkojad? 3. Jaki jest miesiąc roku? 4. Jak mam na imię? 5. Dave jest starszy od Steve'a, ale Steve jest starszy od Jane. Kto jest młodszy, Steve, czy Jane? 6. Jaki dzień będzie jutro? 7. Jakie jest Twoje ulubione jedzenie? 8. Mój przyjaciel Chris lubi grać w futbol. W jakie gry Ty lubisz grać? 9. Jesteś człowiekiem czy komputerem? 10. Jaka litera jest po T. 11. Jak ma na imię mój przyjaciel, który lubi grać w futbol? 12. Jak masz na imię? 13. Co mogę zrobić ze śrubokrętem? 14. Ile liter jest w słowie 'banan'? 15. Czy widziałeś ostatnio dobry film? 16. Jaki rok będzie w przyszłym roku? 17. Jaki jest Twój ulubiony owoc? 18. Wolisz białą czy czarną kawę? 19. Ile masz lat? 20. Piłka została kopnięta przez Freda. Kto kopnął piłkę?”, [tłumaczenie własne]

Źródło: <http://loebner.exeter.ac.uk/selection-results/>, dostęp 30.10.2013

How are you doing today? I didn't hear you. I didn't hear you.

I am not speaking – I am typing

OK. Do you have any pets? I didn't hear you. I didn't hear you.

I will ask the questions.

Thank you for answering me. Do you play any musical instruments?

I am sure you are a pretty dumb program – do you agree?

I happen to like computers a lot. I owe my life to them. How old are you?"

(„Czy mógłbyś kogoś zabić dźgając go rącznikiem?

Jak się masz? Nie słyszałem Cię. Nie słyszałem Cię.

Ja nie mówię - ja piszę.

OK. Czy masz jakieś zwierzęta? Nie słyszałem Cię. Nie słyszałem Cię.

Ja będę zadawać pytania.

Dziękuję za udzielenie odpowiedzi. Czy grasz na jakimś instrumentach?

Jestem pewien, że jesteś całkiem głupim programem - zgadzasz się ze mną?

Tak się składa, że bardzo lubię komputery. Zawdzięczam im życie. Ile masz lat?"⁸²)

Jak widać, poziom finalistów konkursu nie zapowiada na razie rychłego przyznania ani złotego, ani nawet srebrnego spośród medali przewidzianych dla zwycięzców. Konkurs, częściowo ze względu na niekoniecznie zadowalające rezultaty laureatów, jest też często krytykowany jako czynnik nie tylko nie przyspieszający, ale wręcz spowalniający rozwój sztucznej inteligencji jako dziedziny nauki. Jednym z głównych przeciwników konkursu jest Marvin Minsky, który zasłynął nawet ufundowaniem własnej nagrody, w wysokości stu dolarów, dla pierwszej osoby, która odwiedzi Loebnera od dalszego organizowania rzeczono konkursu.⁸³

⁸² [tłumaczenie własne]

Noel Sharkey, *Would artificial intelligence outsmart me? I needn't have worried.* [online] [dostęp: 28.10.2013]. Dostępny: <http://www.theguardian.com/commentisfree/2011/oct/25/artificial-intelligence-loebner-prize>,

⁸³ Znany jest komentarz Minsky'ego do ogłoszenia o konkursie Loebnera na grupie dyskusyjnej comp.ai i comp.ai.philosophy (ogłoszenie zawierało na końcu groźbę, że jeżeli nazwy Loebner Prize i Loebner Prize Competition będą bezprawnie używane, to Loebner wycofa się z organizacji konkursu):

"I do hope that someone will volunteer to violate this proscription so that Mr. Loebner will indeed revoke his stupid prize, save himself some money, and spare us the horror of this obnoxious and unproductive annual publicity campaign.

In fact, I hereby offer the \$100.00 Minsky prize to the first person who gets Loebner to do this. I will explain the details of the rules for the new prize as soon as it is awarded, except that, in the meantime, anyone is free to use the name "Minsky Loebner Prize Revocation Prize" in any advertising they like, without any licensing fee."

W tej chwili środowisko naukowe odnosi się do nagrody Loebnera dość sceptycznie. Jednak nie zawsze tak było – na początku funkcjonowania konkursu w składzie komitetu organizacyjnego byli Daniel Dennet, Joseph Weizenbaum i szereg innych uznanych naukowców, a sponsorami pierwszej jego edycji były rządowa National Science Foundation i Sloan Foundation.

Podsumowanie:

Trudno ocenić wartość nagrody Loebnera dla sztucznej inteligencji. Część krytyki konkursu wynika z zarzutów wysuwanych przeciwko samemu testowi Turinga⁸⁴, część dotyczy formuły konkursu. Mimo krytyki środowiska naukowego⁸⁵ konkurs ten odbywa się rokrocznie i cieszy się dość dużym zainteresowaniem mediów, jako jedyna formalna adaptacja testu Turinga.

Debata wokół zasadności samego Testu Turinga nie ustaje praktycznie od momentu jego powstania. Jednak nawet jeśli zgodzić się z krytykami TT i uznać pomysł Turinga za wadliwy, to nadal można przyznać rację Wiliamowi Rapaportowi:

„Nevertheless, over 50 years after its appearance, the Turing Test continues to help focus debate on issues concerning the relation of cognition to computation.”

(“Niemniej jednak, po 50 latach po jego pojawieniu się, Test Turinga cały czas pomaga w skupieniu debaty na zagadnieniach związanych z relacją poznania do obliczeń komputerowych”)⁸⁶

(„Naprawdę mam nadzieję, że ktoś zaofiaruje się by złamać ten zakaz, żeby Pan Loebner rzeczywiście odwołał swoją głupią nagrodę, zachował swoje pieniądze i oszczędził nam horroru tej przykrej i bezproduktywnej dorocznej kampanii reklamowej.

Istotnie, niniejszym oferuję 100.00\$ na nagrodę Minsky’ego pierwszej osobie, której uda się skłonić Loebnera, żeby to zrobił. Wyjaśnię szczegóły zasad nowej nagrody kiedy tylko zostanie przyznana, poza tym, że w międzyczasie, każdy może używać nazwy „Nagroda Minky’ego Wycofania Nagrody Loebnera” we wszystkich reklamach, w których chce, bez żadnych opłat licencyjnych.”), [tłumaczenie własne], Źródło: <http://loebner.net/Prizef/minsky.txt>, dostęp: 28.10.2013

⁸⁴ Mówi się, że więcej wysiłku włożono w próby zdezawuowania testu Turinga niż w próby jego zdania. Niestety krytyka TT jest tematem na tyle szerokim, że nie można jej tutaj przytoczyć.

⁸⁵ Poza Marvinem Minsky’em konkurs w obecnej formie krytykował również Daniel Dennet oraz Stuart Shieber

⁸⁶ William J. Rapaport, *The Turing Test*

IV.2. SZTUKA PODSTĘPU⁸⁷

"It is said that to explain is to explain away. This maxim is nowhere so well fulfilled as in the area of computer programming, especially in what is called heuristic programming and artificial intelligence. For in those realms machines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away; it stands revealed as a mere collection of procedures, each quite comprehensible."⁸⁸

Joseph Weizenbaum

IV.2.1. ELIZA

Nawet jeżeli moment, w którym jakiemuś programowi lub maszynie uda się spełnić warunki postawione przez Turinga, jest jeszcze bardzo odległy, to nie można odmówić sztucznej inteligencji pewnych sukcesów w dziedzinie imitacji ludzi. Jednym z najsłynniejszych przykładów aplikacji symulujących rozmowę jest ELIZA, stworzona przez Josepha Weizenbaum'a. Program ten, powstały już w 1966 roku, był żartobliwą próbą imitowania terapeuty ze szkoły psychoterapii

⁸⁷ Tytuł podrozdziału zapożyczony z tytułu książki Kevina Mitnicka (w oryginale *The Art of Deception: Controlling the Human Element of Security*)

⁸⁸ "Mówi się, że znaleźć wytłumaczenie to znaleźć usprawiedliwienie. Ta maksyma nigdzie nie sprawdza się tak, jak w dziedzinie programowania, a szczególnie w tym, co nazywamy programowaniem heurystycznym i sztuczną inteligencją. Ponieważ maszyny są stworzone, żeby zachowywać się w tych obszarach w zdumiewający sposób, często wystarczający by olśnić nawet najbardziej doświadczonych obserwatorów. Ale kiedy konkretny program zostanie zdemaskowany, kiedy jego wewnętrzne mechanizmy zostaną wyjaśnione przy pomocy dostatecznie przejrzystego języka, by wywołać zrozumienie, czar pryska; zostaje ujawniony jako zwykły zestaw procedur, wszystkich całkiem zrozumiałych." [tłumaczenie własne]

Joseph Weizenbaum, *ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine*, "Communications of the ACM", 1966, t. 9, nr 1

humanistycznej⁸⁹. ELIZA była oparta na dopasowywaniu wzorców, czyli formułowała odpowiedzi na zasadzie wyłapywania słów i fraz kluczowych, oraz wykorzystywaniu przypisanych do nich „canned responses”⁹⁰ i fragmentów wypowiedzi użytkownika (stosownie zmieniając w nich zaimki osobowe, dzierżawcze i, w bardzo ograniczonym zakresie, odmianę, np. z „I am” na „You are”) do formułowania własnych wypowiedzi.

Przykładowo, jeżeli zadamy Elizie pytanie:

“Do you remember your childhood?”

To dokona ona, w ramach skanowania tekstu, automatycznych zamian (YOU=I) i (YOUR=MY), otrzymując w efekcie zdanie „DO I REMEMBER MY CHILDHOOD”. Kolejnym zadaniem jest znalezienie odpowiedniej dekompozycji wpisanego tekstu, czyli innymi słowy, znalezienie najlepiej dopasowanego szablonu dla zadanego zdania.

W tym przypadku wprowadzone pytanie sprowadza się do postaci (0 DO I REMEMBER 0) gdzie zero oznacza dowolną ilość znaków. Kiedy program znajdzie już odpowiednią regułę dekompozycji zdania, skorzysta ze stosownej reguły transformacji dla danej frazy kluczowej, w tym przypadku:

- ((0 DO I REMEMBER 0) (DID YOU THINK I WOULD FORGET 5)
(WHY DO YOU THINK I SHOULD RECALL 5 NOW)
(WHAT ABOUT 5) (=WHAT) (YOU MENTIONED 5))

Oznacza to, że program wybierze jeden z szablonów odpowiedzi, podstawiając pod 5 piąty komponent zdania wejściowego (1=”” 2=DO 3=I 4=REMEMBER 5=YOUR CHILDHOOD) i tym samym generując ostateczną odpowiedź, np. „Did you think I would forget my childhood?”

Jeżeli wybrana zostanie pozycja (=WHAT) program zastosuje szablon dla frazy kluczowej WHAT:

⁸⁹ A konkretnie rogeriańskiej, stworzonej przez Carla Rogersa, jednego z głównych przedstawicieli psychologii humanistycznej. Eliza mogła, przy pomocy skryptów, imitować różne postaci, ale swoje najbardziej znane wcielenie zawdzięcza skrypcowi DOKTOR, czyli właśnie symulacji psychoterapeuty.

⁹⁰ Ang. „gotowe odpowiedzi”, dosłownie „odpowiedzi z puszki”

- (WHAT ((O) (WHY DO YOU ASK) (DOES THAT QUESTION INTEREST YOU) (WHAT IS IT YOU REALLY WANT TO KNOW)[...])⁹¹

I postąpi analogicznie jak w poprzednim przypadku. Oznacza to, że program na podane pytanie udzieli jednej z następujących odpowiedzi:

Did you think I would forget my childhood?
 Why do you think I should recall my childhood now ?
 Why do you think I should recall my childhood now?
 What about my childhood?
 You mentioned my childhood?

Lub, gdy wybrana zostanie czwarta pozycja, np:

Why do you ask?

Przy czym wybór kolejnego szablonu odpowiedzi zależy od tego po raz który program odnotował wystąpienie danej frazy kluczowej. W związku z tym, jeżeli zadamy Elizie pytanie "Do you remember your childhood?" i będzie to pierwsze dopasowanie dekompozycji (0 DO I REMEMBER 0), to z całą pewnością otrzymamy odpowiedź „Did you think I would forget my childhood?”.

Jak widać w powyższym przykładzie, ELIZA może, przy pomocy sprytnie opracowanych reguł, udzielać dość zróżnicowanych (choć powtarzających się) odpowiedzi, składających się na iluzję zatroskanego psychoterapeuty. Temat psychoterapii nie został bynajmniej wybrany przypadkowo. W związku z charakterystyką typowego wstępnego wywiadu psychologicznego, można było stosunkowo łatwo ustalić skończony zbiór sensownie brzmiących wypowiedzi programu.

"This mode of conversation was chosen because the psychiatric interview is one of the few examples of categorized dyadic natural language communication in which

⁹¹ Fragment oryginalnego skryptu DOKTOR

one of the participating pair is free to assume the pose of knowing almost nothing of the real world."

(“Ten tryb konwersacji został wybrany ze względu na to, że wywiad psychiatryczny jest jednym z niewielu przykładów usystematyzowanej dwustronnej komunikacji w języku naturalnym, w której jedna ze stron może przyjąć postawę kogoś, kto nie wie niemal niczego o świecie rzeczywistym.”)⁹²

W oczywisty sposób ułatwia to pracę twórcy programu, który nie musi dzięki temu wyposażać swojej aplikacji w żadną wiedzę o świecie. Zamiast tego może bazować na utrzymywaniu iluzji, w jakiej pogrążony jest rozmówca programu (od początku poinformowany przecież o tym, że rozmawia przez terminal z prawdziwym terapeutą) przy pomocy stosowania subtelnych sztuczek psychologicznych (zdarza się, że kiedy ludzie używają słów takich jak „zawsze”, „nikt” i „wszyscy” w zdaniach takich jak „wszyscy się ze mnie śmieją”, w rzeczywistości mają na myśli kogoś konkretnego, dlatego też jeżeli Eliza wyłapie któreś z tych słów w wypowiedzi użytkownika, to odpowie pytaniem o to czy jednak nie mamy na myśli kogoś konkretnego, wzmagając tym samym u rozmówcy iluzję, że maszyna go rozumie) oraz odpowiedzi w rodzaju:

X: I like watching films.

Eliza: Do you say you like watching films for some special reason?

Gdyby rozmowa ta nie miała odbywać się między terapeutą a pacjentem, a np. na internetowym chacie, to naturalną odpowiedzią Elizy musiałaby być jakaś uwaga na temat jej własnych preferencji odnośnie filmów lub na temat filmów w ogóle, zdradzająca jakąkolwiek wiedzę Elizy na ten temat.

Pomimo niewielkiego zbioru reguł, Elizie udawało się wprowadzać w błąd ludzi, którzy sądzili, że faktycznie rozmawiają z prawdziwym terapeutą i zwierzali się programowi ze swoich problemów. Co ciekawe, Weizenbaum miał nawet trudności z przekonaniem niektórych rozmówców swojego programu, że Eliza naprawdę *nie jest* człowiekiem.

⁹² [tłumaczenie własne], J. Weizenbaum

Kolejnym znanym przykładem wczesnego, bo powstałego już w 1972 roku, chatbota⁹³ (czyli właśnie programu symulującego rozmowę) jest PARRY autorstwa Kenneth'a Mark'a Colby'ego, który miał za zadanie udawać psychicznie chorego pacjenta. Poniżej przytaczam fragment rozmowy Elizy z Parry'm, czyli pogawędki sztucznego terapeuty z symulacją paranoika:

Parry: You harp on the same questions too much.

Eliza: What does that similarity suggest to you?

Parry: I don't get you.

Eliza: Do you wish that you did get it?

Parry: It's about time you showed some interest in my feelings.

Eliza: Oh, I showed some interest in your feelings?"⁹⁴

(„**Parry:** Zbyt dużo gładzisz o tych samych pytaniach.

Eliza: Co sugeruje Ci to podobieństwo?

Parry: Nie rozumiem Cię.

Eliza: Czy chciałbyś mnie rozumieć?

Parry: Najwyższy czas, żebyś wykazała jakieś zainteresowanie moimi uczuciami.

Eliza: Oh, wykazałam jakieś zainteresowanie Twoimi uczuciami?"

Trzeba przyznać, że jest to zaskakująco ludzka konwersacja, jak na dwa programy powstałe odpowiednio 47 i 41 lat temu. Mogłoby się wydawać, że skoro stosunkowo proste zasady pozwoliły na osiągnięcie dobrych rezultatów na tak wczesnym etapie rozwoju komputeryzacji, to do roku 2013 będzie można porozmawiać z robotem praktycznie na każdy temat. Jednak jak się okazuje, nic bardziej mylnego. Co więcej, ELIZA stanowi nadal dobry punkt wyjścia dla opisu

⁹³ Spotykana jest również nazwa chatterbot.

⁹⁴ [tłumaczenie własne]

Zapis rozmowy Elizy z Parrym można znaleźć w książce George'a Johnsona *The Machinery of the Mind*.

dużej części programów konwersacyjnych, co będzie widoczne w następnym podrozdziale.

IV.2.2. CHATBOTY

IV.2.2.1. A.L.I.C.E. i AIML

Trudno stwierdzić, czy da się wskazać jeden najlepszy sposób programowania chatbotów, ale z całą pewnością można łatwo wskazać technologię, która zdominowała dziedzinę tworzenia programów symulujących konwersację, przynajmniej pod względem popularności - jest nią język znaczników stworzony przez Richarda S. Wallace'a. Mowa tutaj o AIML'u, czyli Artificial Intelligence Markup Language, dialekcie XML'a, służącym do tworzenia chatbotów bazujących na rozpoznawaniu wzorców. Na AIML'u oparta jest spora część botów startujących w konkursie o nagrodę Loebnera⁹⁵, jak również, a może i przede wszystkim 221,000 botów⁹⁶ na pandorabots.com, serwisie umożliwiającym darmowe stworzenie własnego bota i umieszczenie go w sieci.

AIML zawdzięcza swoją popularność temu, że pozwala na stosunkowo łatwe i szybkie tworzenie dużych baz reguł definiujących odpowiedzi chatbotów - teoretycznie, stworzenie własnej aplikacji do konwersacji jest możliwe nawet przy szczątkowej tylko znajomości programowania⁹⁷. Nie ma też konieczności tworzenia wszystkich szablonów od nowa - wystarczy skorzystać z udostępnionych na licencji open source plików użytych przy tworzeniu bota A.L.I.C.E. (którego omówienie znajduje się w dalszej części rozdziału) i rozwijanie własnego programu na ich podstawie.

⁹⁵ Jak chociażby tegoroczny zwycięzca, bot Mitsuku, autorstwa Steve Worswick'a.

⁹⁶ Dane z lutego 2013 roku.

⁹⁷ Szczególnie, jeżeli korzysta się z gotowego silnika dla chatbotów, takiego jak np. program-o, czy też tworzy się własnego chatbota korzystając ze, wspomnianego już, serwisu pandorabots.com

AIML – przegląd możliwości

Podstawową jednostką w AIML’u jest kategoria (*category*). Kategorie zawierają, w swojej podstawowej wersji, jedynie wzorzec (*pattern*) i szablon odpowiedzi (*template*). Przykładowo, aby sprawić, żeby chatbot odpowiadał na powitanie „Hello”, wystarczy napisać:

```
<category>
  <pattern>HELLO</pattern>
  <template>
    Hello.
  </template>
</category>
```

(We wzorcach nie stosuje się interpunkcji oraz używa wyłącznie kapitałek, ponieważ do takiej postaci sprowadzają wypowiedzi użytkownika interpretery AIML’a⁹⁸)

W rzeczywistości ludzie bardzo rzadko reagują na jakąkolwiek wypowiedź (włącznie z powitaniem) za każdym razem w taki sam sposób, dlatego też warto przewidzieć różne wersje odpowiedzi na to samo pytanie. Taka różnorodność, która w skrypcie dla Elizy mogłoby zostać zapisana np. jako „((0 HELLO 0)(HI)(HELLO)(HI THERE))” w AIML’u jest realizowana przy pomocy tag’ów <random></random> z umieszczoną pomiędzy nimi listą możliwych odpowiedzi. Każdy element listy jest umieszczony między tagami . Przykładowo:

```
<category>
  <pattern>HELLO</pattern>
  <template>
    <random>
      <li>Hi! </li>
      <li>Hello. </li>
      <li>Hi... </li>
    </random>
  </template>
</category>99
```

Chatbot wyposażony w taki szablon odpowiedzi na powitanie Hello udzieli, losowo, jednej z przewidzianych odpowiedzi „Hi!”, „Hello.” lub „Hi...”.

⁹⁸ W swoim programie wykorzystuję PyAIML, interpreter AIML dla Pythona, stworzony przez Corta Strattona

⁹⁹ Przykłady kodu w tym rozdziale, jeżeli nie zaznaczono inaczej, pochodzą bezpośrednio z załączonego do pracy programu (lub stanowią uproszczoną wersję zastosowanego w programie rozwiązania).

Siła ELIZY polegała w dużej mierze na umiejętnym wykorzystywaniu fragmentów wypowiedzi użytkownika. Podobne efekty można osiągnąć również w AIML'u przy pomocy znaku * umieszczonego we wzorcu, który zastępuje dowolną (niezerową¹⁰⁰) ilość znaków. Odwołanie się do fragmentu zdekomponowanego zdania odbywa się tutaj poprzez tag <star/> umieszczony wewnątrz szablonu. AIML pozwala też na automatyczną zamianę zaimków osobowych i dzierżawczych dzięki tagom <person></person> (zamieniającym pierwszą osobę z trzecią) i <person2></person2>¹⁰¹ (zamieniającym pierwszą osobę z drugą).

```
<category>
  <pattern>
    DO YOU REMEMBER *
  </pattern>
  <template>
    <random>
      <li>Did you think I would forget <person2><star/></person2>?</li>
      <li>Why do you think I should recall <person2><star/></person2>now?</li>
      <li>What about <person2><star/></person2>?</li>
      <li>You mentioned <person2><star/></person2>?</li>
    </random>
  </template>
</category>
```

Istnieje również możliwość odwołania się w jednym szablonie do szablonu drugiego wzorca. To, co w przykładowym fragmencie skryptu Elizy z podrozdziału IV.2.1. można było osiągnąć poprzez użycie w szablonie (=WHAT), w AIML'u realizuje się przy pomocy tagów <srail></srail>, pomiędzy którymi umieszcza się wzorzec, z którego szablonu chcemy skorzystać.

```
<category>
  <pattern>
    DO YOU REMEMBER *
  </pattern>
  <template>
    <random>
      <li>Did you think I would forget <person2><star/></person2>?</li>
      <li>Why do you think I should recall <person2><star/></person2>now?</li>
      <li>What about <person2><star/></person2>?</li>
      <li><srail>WHAT</srail></li>
      <li>You mentioned <person2><star/></person2>?</li>
    </random>
  </template>
</category>
```

¹⁰⁰ Co stanowi jedną z większych wad AIML'a, tworząc konieczność zastępowania konstrukcji (O SOMETHING 0) trzema kategoriami o wzorcach *SOMETHING, SOMETHING* i *SOMETHING*

¹⁰¹ Tagi <person> i <person2> mogą występować tylko wewnątrz szablonu odpowiedzi.


```

<category>
  <pattern>
    WHAT
  </pattern>
  <template>
    <random>
      <li>why do you ask?</li>
    </random>
  </template>
</category>

```

AIML pozwala też na udzielanie odpowiedzi w zależności od tego, jaka była poprzednia wypowiedź programu. Służą do tego tag `<that></that>` umieszczany między znacznikami wzorca a znacznikami szablonu.

```

<category>
  <pattern>
    DO YOU LIKE TEA
  </pattern>
  <template>
    I love it! Especially sencha. Do you like tea?
  </template>
</category>

<category>
  <pattern>
    I DO
  </pattern>
  <that>I LOVE IT ESPECIALLY SENCHA DO YOU LIKE TEA</that>
  <template>
    That's great. We have something in common then.
  </template>
</category>

```

W tym przypadku jeżeli po dialogu:

User: Do you like tea?

Program: I love it! Especially sencha. Do you like tea?

Użytkownik udzielił odpowiedzi "I do.", to program skorzysta z reguły zawartej w znacznikach `<that></that>` i odpowie: „That's great. We have something in common then.". Jeżeli jednak wcześniejsza konwersacja była inna niż przedstawiona wyżej, to wzorzec I DO nie zostanie wybrany jako stosowna dekompozycja zdania wejściowego.

W skryptach dla Elizy zawsze musiał istnieć jeden wzorzec, który był wykorzystywany, jeżeli nie udało się dopasować frazy na wejściu do żadnego innego wzorca. AIML oferuje rozwiązanie pozwalające na tworzenie wielu takich wzorców w zależności od aktualnego tematu rozmowy. Można to osiągnąć przy pomocy ustalenia tematu `<set name="topic">NAZWA_TEMATU</set>`.

```
<category>
  <pattern>*</pattern>
  <that>DO YOU LIKE FAMILY ACTIVITIES</that>
  <template>
    I like to play board games with my family. What are your hobbies?
  <think>
    <set name="topic">HOBBY</set>
    <set name="topic_hobby">VISITED</set>
  </think>
</template>
</category>
```

```
<topic name="HOBBY">
```

```
<category>
  <pattern>*</pattern>
  <template>
    <random>
      <li>Do you collect anything?</li>
      <li>Are you a geek? In a positive sense? :)</li>
      <li>Do you think that having a hobby is important in your life?</li>
    </random>
  </template>
</category>
```

```
...
```

```
</topic>
```

(Tag `<think></think>` jest umieszczany wewnątrz szablonu i otacza fragment reakcji programu, który nie ma być wyświetlony na ekranie)

W powyższym przykładzie po ustaleniu tematu na hobby, program będzie szukał dopasowania wzorca następnej wypowiedzi użytkownika tylko pomiędzy tagami `<topic>`, aż temat nie zostanie powtórnie zmieniony. Jeżeli program napotka wypowiedź dla której nie znajdzie innego wzorca w ramach tematu, to odwoła się do wzorca ogólnego `<pattern>*</pattern>` i wybierze jedną z losowych odpowiedzi, np. „Do you collect anything?”

Podobnie jak nazwy tematu, możemy w AIML’u definiować także inne zmienne według własnego uznania, co pozwala na zapamiętanie np. imienia użytkownika

lub, jak wyżej, czy dany temat był już poruszany (<set name="topic_hobby">VISITED</set>). Informacje te można potem odczytywać (<get name="nazwa"/>) i wykorzystywać jako część odpowiedzi lub w tagach <condition>, pozwalających na zróżnicowanie reakcji programu w zależności od różnych zmiennych.

```
<category>
  <pattern>
    WHAT IS MY AGE
  </pattern>
  <template>
    <condition name="userage_status" value="KNOWN">
      I'm not sclerotic!
      You are <get name="userage"/>.
    </condition>
    <condition name="userage_status" value="UNKNOWN">
      You didn't tell me.
    </condition>
  </template>
</category>
```

W powyższym przykładzie odpowiedź programu na pytanie „What is my age?” jest uzależnione od tego, czy użytkownik udzielił wcześniej informacji o swoim wieku.

Przedstawiony wyżej przegląd możliwości AIML’a nie jest wyczerpujący, ale ujawnia pewne słabości tej technologii, częściowo tożsame ze słabościami programów z lat sześćdziesiątych (takie jak konieczność tworzenia ogromnych ilości kategorii dla stworzenia iluzji sensownej rozmowy¹⁰², nienajlepsza czytelność kodu, brak zaawansowanej kontroli nad zmiennymi, nierozróżnianie części mowy oraz pytań od twierdzeń¹⁰³, „sztywność” wzorców itp.) które można zaobserwować również podczas analizy kategorii najbardziej znanego bota opartego na tej technologii, czyli A.L.I.C.E.

A.L.I.C.E. (The Artificial Linguistic Internet Computer Entity)¹⁰⁴

ALICE jest trzykrotnym laureatem brązowego medalu w konkursie o nagrodę Loebnera, a jej autorem jest Dr Richard S. Wallace, czyli sam twórca omawianego tutaj języka. Na uwagę zasługuje przede wszystkim ogromna „baza wiedzy”

¹⁰² Poza wspomnianym już szczególnym przypadkiem rozmowy ściśle zdefiniowanej.

¹⁰³ Ze względu na wspomniany wcześniej fakt, że wzorce nie mogą zawierać znaków interpunkcyjnych.

¹⁰⁴ Sztuczna Lingwistyczna Internetowa Jednostka Komputerowa

A.L.I.C.E., składająca się z około 41 tysięcy (!) kategorii. Poniżej przytaczam fragment zapisu mojej własnej rozmowy z A.L.I.C.E.¹⁰⁵:

judge: What's your name?

ALICE: ALICE.

judge: Why is it written in capital letters?

ALICE: Why do I exist?

judge: I don't know, why do you exist?

ALICE: Good metaphysical question.

judge: Better than some, probably.

ALICE: Perhaps better, but not perfectly.

judge: What is better?

ALICE: You decide.

judge: Do you prefer tea or coffee?

ALICE: Yes Tea is a very good beverage.. Yes I love coffee. It is one of life's necessities. I take it strong.

judge: I prefer tea.

ALICE: What else do you like?

judge: I like reading books

ALICE: You like reading books.

judge: Yes. And you?

ALICE: Wow! I am not sure what you are asking.

(judge: Jak masz na imię?

ALICE: ALICE.

judge: Czemu jest napisane wielkimi literami?

ALICE: Czemu istnieję?

judge: Nie wiem, czemu istniejesz?

ALICE: Dobrze metafizyczne pytanie.

judge: Przypuszczalnie lepsze od niektórych.

ALICE: Być może lepsze ale nie perfekcyjnie.

judge: Co jest lepsze?

¹⁰⁵ Rozmowy z A.L.I.C.E. można przeprowadzać pod adresem http://sheepridge.pandorabots.com/pandora/talk?botid=b69b8d517e345aba&skin=custom_iframe, [dostęp: 4.11.2013]

ALICE: Ty zdecyduj.

judge: Wolisz herbatę czy kawę?

ALICE: Tak Herbata jest bardzo dobrym napojem.. Tak kocham kawę. Jest jedną z konieczności życia. Zamawiam mocną.

judge: Ja wolę herbatę.

ALICE: Co jeszcze lubisz?

judge: Lubię czytać książki

ALICE: Lubisz czytać książki.

judge: Tak. A Ty?

ALICE: Wow! Nie jestem pewna o co pytasz.)

Jak widać, nawet stosunkowo krótka rozmowa demaskuje A.L.I.C.E. jako chatbota. Możliwe, że dopasowywanie wzorców po prostu nie jest metodą, która pozwoli kiedykolwiek na stworzenie symulacji inteligentnej rozmowy. Możliwe też, że A.L.I.C.E., by dobrze funkcjonować, potrzebuje po prostu jeszcze większej ilości kategorii. Z takim stanowiskiem zdecydowanie zgadza się twórca bota, Richard Wallace:

"In the years to come we will only advance the frontier further. [...] The work of those seeking to explain natural language in terms of something more complex than stimulus response will take place beyond our frontier, increasingly in the hinterlands occupied by only the rarest forms of language. [...] Expanding the borders even more we will continue to absorb the stragglers outside, until the very last human critic cannot think of one sentence to "fool" A.L.I.C.E.."

(*"W nadchodzących latach będziemy tylko przesuwać granicę dalej. [...] Praca tych, którzy próbują wyjaśnić język naturalny w kategoriach czegoś bardziej skomplikowanego niż reakcja na bodźce będzie zajmować miejsce poza naszą granicą, coraz bardziej na obrzeżach okupowanych tylko przez najrzadsze formy języka. [...] Poszerzając granice jeszcze dalej będziemy kontynuować pochłanianie maruderów z zewnątrz, dopóki ostatni ludzki krytyk nie będzie był w stanie wymyśleć choćby jednego zdania, by „wykiwać” A.L.I.C.E."*)¹⁰⁶

¹⁰⁶ Wypowiedź pochodzi z wywiadu udzielonego przez Richarda Wallace'a witrynie Slashdot. <http://news.slashdot.org/story/02/07/26/0332225/alicebot-creator-dr-richard-wallace-expounds>, [dostęp: 4.11.2013], [tłumaczenie własne]

IV.2.2.2. *Is Cleverbot clever?*¹⁰⁷

Rozwiązaniem bazującym na dopasowaniu wzorców, ale oferującym pewne bardziej zaawansowane funkcje, takie jak rozróżnianie części mowy, pytań od zdań oznajmujących i generalnie większą łatwość tworzenia wzorców, jest ChatScript autorstwa Bruce'a Willcox'a. Na ChatScriptcie oparty był m.in. zwycięski program konkursu o nagrodę Loebnera z 2010 i 2011 roku, czyli wspomniany już chatbot Rosette.

Przykładowo, wzorce (zebrane w temat) w ChatScriptcie mogą mieć postać:

topic: ~DEATH [dead corpse death die body]

t: I don't want to die

?: (When will you die) I don't know.¹⁰⁸

Zasadniczo odmienne rozwiązanie zostało zastosowane w, święcącym triumfy popularności, Cleverbocie, będącym następcą Jabberwacky'ego, również laureata nagrody Loebnera. Chatbot ten wpisuje się mniej lub bardziej w trend przetwarzania dużych ilości danych zebranych od użytkowników za pośrednictwem Internetu. Każda wypowiedź Cleverbota jest bezpośrednim przytoczeniem wypowiedzi wprowadzonej przez człowieka (obecnie Cleverbot udostępnia też użytkownikom opcje edycji złych odpowiedzi bota). Sztuka polega więc na odpowiednim dopasowaniu którejś z przeszłych wypowiedzi, znajdujących się w bazie, do aktualnego tematu rozmowy. Można powiedzieć, że rozmowa z Cleverbodem jest jak rozmowa z duchami internetowej aktywności. Poniżej przedstawiam fragment własnej rozmowy z Cleverbodem:

User: Tell me something about yourself.

Cleverbot: Ok, I live in New York City, I am 16 years old, I love to read and write and I hope to make a career out of it one day.

User: Who is your favourite writer?

¹⁰⁷ Ang. „Czy Sprytny Bot jest sprytny?”

¹⁰⁸ Przykład pochodzi z dokumentacji, dostępnej na:

<http://sourceforge.net/projects/chatscript/files/>

Bruce Wilcox, *ChatScript Basic User Manual*. [online]. [dostęp: 15.11.2013]. Dostępny:

<http://sourceforge.net/projects/chatscript/files/ChatScript%20Basic%20User%20Manual.pdf/download>

Cleverbot: My favourite color is orange.

User: I prefer green. Why do you like orange?

Cleverbot: I love blue and pink.

User: So which one is your favourite?

Cleverbot: Ferris Bueller's Day Off because Matthew Broderick=awesome

(Użytkownik: Powiedz mi coś o sobie.

Cleverbot: Ok, mieszkam w Nowym Yorku, mam 16 lat, kocham czytać I pisać i mam nadzieję, że kiedyś uda mi się zrobić z tego sposób na życie.

Użytkownik Kto jest Twoim ulubionym pisarzem?

Cleverbot: Mój ulubiony kolor to pomarańczowy.

Użytkownik Wolę zielony. Czemu lubisz pomarańczowy?

Cleverbot: Kocham niebieski i różowy.

Użytkownik: Więc który jest Twój ulubiony?

Cleverbot: Wolny dzień pana Ferrisa Buellera ¹⁰⁹ bo Matthew Broderick=niesamowity)¹¹⁰

Cleverbot może zaskakiwać użytkowników nieformalnymi, a czasem nawet niegrzecznymi wypowiedziami czy też zaskakującymi porównaniami. Jednak jak widać w powyższym przykładzie, jego wypowiedzi często nie pasują zupełnie do kontekstu rozmowy.

IV.2.2.3. Co chatbot wiedzieć i umieć powinien?

Obecnie głównym czynnikiem skłaniającym ludzi do tworzenia chatbotów nie jest wcale chęć wykreowania idealnej symulacji ludzkiego rozmówcy. Chatboty są szeroko stosowane w biznesie (jako tak zwani wirtualni doradcy) i rozrywce (szczególnie w grach komputerowych). Dla większości zastosowań komercyjnych boty bazujące na rozwiązaniach takich jak AIML, czy ChatScript mogą być zupełnie wystarczające. Gdyby jednak chcieć stworzyć chatbota, który sprawnie udawałby człowieka, to w jakie umiejętności wypadałoby go wyposażyć?

¹⁰⁹ „Ferris Bueller's Day Off” to tytuł filmu z 1986 roku. Matthew Broderick był odtwórcą głównej roli.

¹¹⁰ Rozmowy z Cleverbotem można przeprowadzać pod adresem <http://www.cleverbot.com/> [dostęp:5.11.2013]

Turing sugerował stworzenie najpierw umysłu-dziecka, a następnie uczenie go, aż stanie się zdolny do rozmowy na poziomie dorosłego człowieka. Z pewnością chatboty powinny mieć możliwość uczenia się z przeprowadzanych z nimi rozmów i faktycznie, niektóre, jak Cleverbot, stają się tym sprytniejsze, im częściej się z nimi rozmawia. Uczenie z kolei nierozzerwalnie wiąże się z posiadaniem pamięci, a przez to jakiegoś sposobu reprezentacji wiedzy. Weizenbaum sugerował nawet, że programy konwersacyjne powinny tworzyć jakiś rodzaj wewnętrznej reprezentacji swojego rozmówcy (być może powinny tworzyć również wewnętrzną reprezentację rzeczywistości?).

Po inteligentnym rozmówcy możemy też spodziewać się umiejętności logicznego rozumowania, wyciągania wniosków, rozpoznawania kontekstu rozmowy, czy też przeprowadzania prostych obliczeń matematycznych. Do uzupełnienia iluzji brakuje jeszcze, aby wypowiedzi bota świadczyły w jakichś sposób o jego emocjach, by miewał różne nastroje.¹¹¹

Być może taki chatbot musiałby po prostu bardziej umiejętnie stosować „sztuczki” stosowane już teraz, czyli udzielać odpowiedzi sprawiających wrażenie "tajemniczych", zmieniać nieznanne tematy na inne, odpowiadać zagadkami i pytaniami, zapamiętywać niektóre informacje o użytkownikach, rozpoznawać nieistniejące słowa i niezrozumiałe zdania¹¹², czy też stosować naumyślne błędy i opóźnienia.¹¹³

Możliwe też, że wszystkie te rzeczy mogłyby być efektem zastosowania symulacji opartej na którejś z teorii umysłu. Dzisiaj bardziej prawdopodobne jest, że

¹¹¹ Większość tych cech próbuje się implementować, jak widać, z różnym skutkiem.

¹¹² Co jest stosunkowo proste do osiągnięcia przy użyciu modeli n-gramów opartych na łańcuchach Markova.

¹¹³ Powyższe założenia dotyczą programów mających za zadanie zdać test Turinga, a więc porozumiewających się z ludźmi wyłącznie drogą tekstową. Dodatkowe trudności pojawiają się w momencie, w którym konwersacja miałaby odbywać się na zasadzie videorozmowy. Poza oczywistymi trudnościami technicznymi, mogą pojawić się również całkiem niespodziewane problemy. Np. wydawać by się mogło, że przy dzisiejszym stopniu zaawansowania animacji komputerowej stworzenie awatara do rozmowy video, który byłby trudny do odróżnienia od człowieka jest w zasięgu naszych technologicznych możliwości. Okazuje się jednak, że zależność między podobieństwem do oryginału, a pozytywnym odbiorem symulacji (w tym robota) nie jest liniowa. Avatar, który byłby niemal taki sam jak człowiek zostałby określony jako dziwny i odpychający i tym samym zdemaskowany. Odbiór poprawiłby się dopiero w momencie, w którym naśladownictwo zaczęłoby zbliżać się do ideału. Zjawisko to nosi nazwę Uncanny Valley.

eksplorowane będą raczej metody przetwarzania języka naturalnego przy pomocy metod statystycznych, które mogą okazać się skuteczne (dziś święcą triumfy m.in. w dziedzinie tłumaczeń). Jednak jak stwierdził Noam Chomsky, słynny amerykański lingwista i kognitywista, „the »new AI« - focused on using statistical learning techniques to better mine and predict data – is unlikely to yield the explanatory insight that science ought to offer”¹¹⁴

IV.2.2.3. Podsumowanie

Mimo upływu czasu i rozwoju technologii, działanie niektórych chatbotów opiera się na zasadach łądząco podobnych do tych, zgodnie z którymi funkcjonował program Weizenbaum’a. Programy te są jednak pozbawione podstawowej zalety Elizy, a mianowicie elegancji wynikającej z prostoty i efektywności zaproponowanego przez Weizenbaum’a rozwiązania.

Można powiedzieć, że programy te bazują w dużej mierze na trikach i w najlepszym przypadku mogą być posądzone o spryt, ale z pewnością nie o inteligencję. Czas pokaże, które metody (jeżeli którekolwiek) okażą się najbardziej skuteczne i czy doprowadzą one do stworzenia programu symulującego inteligencję, czy też taką inteligencję faktycznie posiadającego.

¹¹⁴ „nowa AI – zorientowana na użycie statystycznych metod uczenia do lepszego przekopywania i przewidywania informacji – nie ma dużych szans na dostarczenie objaśniającego wglądu, jaki powinna oferować nauka.” [tłumaczenie własne]

Cytat pochodzi z wywiadu opublikowanego w *The Atlantic* dostępnego pod adresem: <http://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/v>, [dostęp: 4.11.2013]

PODSUMOWANIE:

Pomimo upływu lat i niezwykle szybkiego rozwoju technologii odpowiedź na pytanie o to czy nadchodzi osobliwość technologiczna jest tak samo niejednoznaczna jak na pytanie czy maszyny mogą posiadać świadomość. Jedyne, co jesteśmy w stanie stwierdzić z całą pewnością to to, że silna sztuczna inteligencja jeszcze nie istnieje (albo, że nie posiadamy odpowiedniego narzędzia diagnostycznego żeby to istnienie stwierdzić).

Test Turinga w swojej współczesnej wersji służy bardziej rozrywce niż nauce i generuje dość rozczarowujące rezultaty i umiarkowanie interesujące technologie. TT ma jednak, moim zdaniem, przynajmniej jedną zasługę – podtrzymuje zainteresowanie pytaniem „Czy maszyny mogą myśleć?”, które z kolei, nawet jeśli samo w sobie ciekawe, ale niekoniecznie praktyczne, pozwala na skupieniu uwagi badaczy i społeczeństwa na etycznych aspektach rozwoju sztucznej inteligencji. Jest też motorem badań nad umysłem i pytań o relacje maszyn i ludzi.

Podsumowując, okazuje się, że zdanie zamykające artykuł Turinga *Computing Machinery and Intelligence* jest tak samo aktualne dziś, jak i w chwili, w którym zostało napisane:

„We can see only a short distance ahead, but we can see that much remains to be done.”¹¹⁵

(„Widzimy tylko na niewielką odległość w przód, ale możemy zobaczyć, że jeszcze wiele pozostało do zrobienia”)

¹¹⁵ A.Turing

BIBLIOGRAFIA:

1. Asimov, Isaac, Lustrzane odbicie [w:] Świat robotów, przeł. Edward Szmigiel, Varia APD, 1993
2. Asimov, Isaac, Roboty i Imperium, przeł. Paulina Braiter, Wyd. 2 popr., Poznań: Dom Wydawniczy Rebis, 2003
3. Bruce Wilcox, ChatScript Basic User Manual.[online].[dostęp: 15.11.2013].
Dostępny:
<http://sourceforge.net/projects/chatscript/files/ChatScript%20Basic%20User%20Manual.pdf/download>
4. Byron, Ada Augusta, notatki do Sketch of The Analytical Engine Invented by Charles Babbage L. F. Menabrei [online].[dostęp: 15.10.2013].Dostępny:
<http://psychclassics.yorku.ca/Lovelace/lovelace.htm>
5. Chomsky, Noam: Noam Chomsky on Where Artificial Intelligence Went Wrong, Rozm. Przepr. Yarden Katz, the Atlantic.[online].[dostęp: 4.11.2013]
Dostępny:
<http://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/v>
6. Clive Thompson, What Is I.B.M.'s Watson?.[online].[dostęp: 7.02.2013].
Dostępny:
<http://www.nytimes.com/2010/06/20/magazine/20Computert.html?page-wanted=1&r=3&>
7. Dennett , Daniel, Cognitive Wheels: The Frame Problem of AI. W: Minds, Machines, and Evolution. Red. C. Hookway. Cambridge University Press, 1984, s. 128-151
8. Dijkstra , E.W., The threats to computing science - ACM South Central Regional Conference [online]. [dostęp: 10.10.2013]. Dostępny:
<http://www.cs.utexas.edu/~EWD/transcriptions/EWD08xx/EWD898.html>
9. Eric Margolis, Richard Samuels, Stephen P. Stich (red.), The Oxford Handbook of Philosophy of Cognitive Science, Oxford University Press, 2012
10. Feigenbaum , Edward A, Some Challenges and Grand Challenges for Computational Intelligence, Journal of the ACM, 2003, tom 50, nr 1, s. 32-40
Geoffrey Jefferson, *The Mind of Mechanical Man*, British Medical Journal, 1949, t. 1, nr 4616

11. Ferrucci , David et al., Building Watson: An Overview of the DeepQA Project ,Ai magazine, s. 68
12. Geoffrey Jefferson, The Mind of Mechanical Man, British Medical Journal, 1949, t. 1, nr 4616
13. Good, I.J, The First Ultraintelligent Machine, Advances in Computers, t.6, 1965
14. Penrose, Roger, *Cienie umysłu*, przeł. Piotr Amsterdamski, Poznań : Zysk i Sk-a Wydawnictwo, 2000, ISBN 83-7150-526-4
15. Jackson, Frank, Epiphenomenal Qualia, Philosophical Quarterly, 1982, nr 32
16. Joseph S. Fulda, „Implications of a logical paradox for computer-dispensed justice reconsidered: some key differences between minds and machines”, Artificial Intelligence and Law, 2012, t. 20, nr 3, s 321-333
17. Joseph Weizenbaum, ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine, “Communications of the ACM”, 1966, t, 9, nr 1
18. Joy, Bill, Why The Future Doesn't Need Us. “Wired” nr 8.04, 2000
19. Kurzweil Ray, The Law of Accelerating Returns, [online]. [dostęp: 15.10.2013]. Dostępny: <http://www.kurzweilai.net/the-law-of-accelerating-returns>
20. Minsky, Marvin Lee, Computation: Finite and Infinite Machines, New Jersey: Prentice-Hall, 1967 Catholic Answers [online].[dostęp:13.10.2013].
21. Minsky, Marvin Lee, The Emotion Machine. Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind, Nowy York: SIMON&SCHUSTER PAPERBACKS, 2006, s. 33, ISBN-13:978-0-7432-7663-4
22. Moravec, Hans, Rise of the Robots, Scientific American, 1999, s.124-135
23. Moravec, Hans, Robots, Re-Evolving Mind at 107 Times Nature's Speed, Cerebrum [online]. 2001, nr 2 [dostęp: 20.10.2013]. Dostępny: <http://www.dana.org/news/cerebrum/detail.aspx?id=3010>
24. Moravec, Hans: Superhumanism. Rozm. Przepr. Charles Platt. Wired, nr 3.10, 1995Simon Herbert, *The Shape of Automation for Man and Management*, New York: Harper&Row, 1965

25. Muehlhauser, Luke, Salamon, Anna, Intelligence Explosion: Evidence and Import [online]. [dostęp: 7.02.2013]. Dostępny: <http://intelligence.org/files/IE-EI.pdf>
26. Penrose, Roger, Cienie umysłu, przeł. Piotr Amsterdamski, Poznań : Zysk i Sk-a Wydawnictwo, 2000, ISBN 83-7150-526-4
27. Phil LeBeau, "Nissan to build self-driven cars".[online].[dostęp:20.10.2013].Dostępny: <http://www.cnbc.com/id/100991243>
28. Sharkey, Noel, Would artificial intelligence outsmart me? I needn't have worried.[online][dostęp: 28.10.2013].Dostępny:<http://www.theguardian.com/commentisfree/2011/oct/25/artificial-intelligence-loebner-prize>
29. Simon Herbert, The Shape of Automation for Man and Management, New York: Harper&Row, 1965
30. Stuart Russel, Peter Norvig (red.), Artificial Intelligence: A Modern Approach, New Jersey: Pearson Education, 2010, ISBN-13: 987-0-13-207148-2
31. Turing, Alan, Computing Machinery and Intelligence [online]. [dostęp: 5.10.2013]. Dostępny: http://www.turingarchive.org/browse.php/B/9Oficjalna_strona_nagrody_Loebnera. [online].[dostęp:30.10.2013].Dostępny:<http://www.loebner.net/Prizef/loebner-prize.html>
32. Waldrop , M. Mitchell, Computer modelling: Brain in a box, Nature, 2012, t. 482 nr 7386, s. 456–458
33. A.L.I.C.E..[online][dostęp: 4.11.2013]. Dostępny:http://sheepridge.pandorabots.com/pandora/talk?botid=b69b8d517e345aba&skin=custom_iframe
34. Cleverbot..[online].[dostęp:5.11.2013]. Dostępny: <http://www.cleverbot.com/>
35. Oficjalna strona nagrody Loebnera.[online].[dostęp:30.10.2013].Dostępny:<http://www.loebner.net/Prizef/loebner-prize.html>

36. Raport: "Self-driving cars: The next revolution".[online].[dostęp 27.10.2013].Dostępny:
<http://www.kpmg.com/Ca/en/IssuesAndInsights/ArticlesPublications/Documents/self-driving-cars-next-revolution.pdf>
37. Strona internetowa teleturnieju Jeopardy.[online].[dostęp: 7.02.2013].
Dostępny: <http://www.jeopardy.com/gamesandmobile/clueoftheday/>
38. Wyniki konkursu o nagrodę Loebnera za rok 2011.[online].[dostęp:30.10.2013]. Dostępny: Źródło:
<http://loebner.exeter.ac.uk/selection-results/>
39. Wypowiedź użytkownika Scott Walker. Dostępny:
<http://forums.catholic.com/showthread.php?t=106016&page=15>
40. Wystąpienie Johna Searle'a w ramach TEDxCERN.[online].[dostęp: 12.10.2013], Dostępny:
http://www.ted.com/talks/john_searle_our_shared_condition_consciousness.html
41. Wystąpienie Martina Rees'a w ramach konferencji TEDGlobal2005.[online].[dostęp:20.10.2013].Dostępny:
http://www.ted.com/talks/martin_rees_asks_is_this_our_final_century.html
42. Wywiad z Richardem Wallace'm w witrynie Slashdot.[online]. [dostęp: 4.11.2013].Dostępny:
[http://news.slashdot.org/story/02/07/26/0332225/alicebot-creator-dr-richard-wallace-expounds,](http://news.slashdot.org/story/02/07/26/0332225/alicebot-creator-dr-richard-wallace-expounds)

SPIS ILUSTRACJI:

ILUSTRACJA I - 1: IDEA TRENINGU PAROWANEGO	10
ILUSTRACJA I - 2: PROCES PRZYSWAJANIA INFORMACJI PRZEZ NELL'A.....	10
ILUSTRACJA I - 3: PRZYKŁAD DZIAŁANIA WOLPHRAM ALPHA	11
ILUSTRACJA I - 4: ROBOT ASIMO.....	14
ILUSTRACJA II - 1: OKŁADKA MAGAZYNU TIME. „CAN MACHINES THINK? THEY ALREADY DO, SAY SCIENTISTIS. [...]”	16
ILUSTRACJA III - 1: OKŁADKA MAGAZYNU WIRED, „WHY THE FUTURE DOESN'T NEED US BY BILL JOY”	41
ILUSTRACJA III - 2 AUTO GOOGLE	50
ILUSTRACJA III - 3 MOTYW UTRATY PRACY W KULTURZE POPULARNEJ.	53
ILUSTRACJA IV - 1: MEDAL DLA ZWYCIĘZCY GŁÓWNEJ NAGRODY KONKURSU O NAGRODĘ LOEBNERA	56
ILUSTRACJA IV - 2: TEST TURINGA.	58
ILUSTRACJA IV - 3: PIERWOTNY KSZTAŁT IMITATION GAME..	58