Paweł Gburzyński

Czy sztuczna inteligencja jest prawdziwa?

Kwestia sztucznej inteligencji pojawiła się na radarze naszej cywilizacji na długo przed narodzeniem się koncepcji współczesnego komputera. Można śmiało zacząć, że już starożytni Grecy fantazjowali na temat mechanicznych tworów obdarzonych zdolnością myślenia.[1] Średniowieczne teorie alchemii określały pojęcie homunkulusa, czyli sztucznego człowieka, odróżniającego się w (raczej drobnych) anatomicznych detalach od ludzkiego wzorca, lecz wyposażonego w zdolność myślenia. Homunkulus pojawia się w Fauście,[2] stworzony przez zdolnego ucznia Fausta – Wagnera, przejawiając wysoką, nadludzką inteligencję oraz tęsknotę stania się autentycznym człowiekiem. Od samego początku wyobrażeń sztucznego rozumu, czy to prokurowanego mechanicznie, jak u Karela Capka,[3] czy z półproduktów biologicznych (lub organicznych), jak u Goethego lub Mary Shelley,[4] zaznacza się dążenie do dorównania człowiekowi lub pokonania go, nie tylko w zakresie chłodnych kompetencji intelektualnych, lecz także w sferze zmysłów i emocji.

Rysunek 1. Homunkulus Fausta (artysta nieznany, Wikimedia Commons).

Wynalazek współczesnego uniwersalnego komputera, za co uznamy propozycję Johna von Neumanna, by algorytm oraz dane, na których on operuje przechowywać z grubsza w ten sam sposób,[5] wytyczył nowy kierunek prac nad homunkulusem, przynajmniej w zakresie podstawowej technologii. Kierunek ten obowiązuje do dziś i zaczyna właśnie owocować produktami, które u jednych budzą zachwyt, u innych zaniepokojenie, a najczęściej jedno i drugie. Współczesny homunkulus zarzucił zamiar

[1] Apollonius Rhodius, *Argonautica, Księga IV*, epizod Talosa, 3w. p.n.e.
[2] Johann Wolfgang von Goethe, *Faust,* Wydawnictwo Siedmioróg, 2017.
[3] Karel Capek, R.U.R. *(Rossum's Universal Robots),* Wildside Press, 2016.w
[4] Mary Shelley, Frankenstein, Wydawnictwo Vesper, 2013.
[5] John von Neumann, *First Draft of a Report on the EDVAC,* Contract No. W–670–ORD–4926 between the United States Army Ordnance Department and the University of Pennsylvania, 1945.

doścignięcia i pokonania człowieka przez próbę emulacji jego zewnętrznej powłoki, która według archaicznych wizji miała stanowić warunek wstępny dla konstrukcji nie-ludzkiego intelektu. Ciałem sztucznej inteligencji, egzemplifikowanej przez ChatGPT,[6] jego kolejne wersje i następniki, jest chmura (albo jak ktoś woli chmara) komputerów o złożoności, o jakiej von Neumannowi śnić się nie mogło. Jej umysłem są głębokie neuronowe modele lingwistyczne wsparte olbrzymim korpusem tekstów wyprodukowanych przez naszą cywilizację.

Słowo „korpus" pojawi się wiele razy w naszej dyskusji. Gdy w roku 1971 rozpoczynałem studia matematyczne na wydziale Matematyki i Mechaniki UW, Michael Stern Hart na University of Illinois wprowadził w komputer Konstytucję Stanów Zjednoczonych inicjując w ten sposób projekt o nazwie Gutenberg,[7] którego celem było zakodowanie wszystkich pisanych materiałów utworzonych przez ludzkość w postać cyfrową – by stały się niezniszczalne i dostępne komputerom. Współczesna sztuczna inteligencja posiada w swej historii wiele kamieni milowych. Dziś mało kto słyszał o tamtym projekcie, ale był to niewątpliwie jeden z nich.

Celem, jaki sobie postawiłem przystępując do pisania niniejszego eseju jest przystępne objaśnienie, co to wszystko znaczy, tak by czytelnik, niezależnie od technicznego przygotowania, nabył po przeczytaniu intuicji dla ustawienia swojej własnej kreski pomiędzy mistyką i rzeczywistością. Warto bowiem podkreślić na wstępie (co bynajmniej nie wydaje się oczywiste na podstawie licznych popularyzatorskich komentarzy, które przeczytałem lub zasłyszałem ostatnio), że sztuczna inteligencja nie jest tworem mistycznym. Jej mechanizmy, aczkolwiek skomplikowane i wymagające zaawansowanej wiedzy technicznej dla pełnego zrozumienia, poddają się rzeczonemu zrozumieniu, gdyż nie zrodziły się przez alchemiczny traf, gdzie akcydentalne zmieszanie kilku przypadkowych substancji, przy drobnej asyście Mefistofelesa, doprowadziło do magicznej realizacji nieznanego przedtem monstrum. W odróżnieniu od faustowskiego homunkulusa, współczesna sztuczna inteligencja stanowi wynik konsekwentnej realizacji pewnego planu technologicznego, który wystartował razem z narodzeniem się pojęcia algorytmu związanego z formą współczesnego komputera. O ile poszczególne odkrycia w każdej dyscyplinie technicznej mogą zależeć od przypadku, funkcjonowanie uzyskanych z ich pomocą konstrukcji jest zawsze do końca zrozumiałe. Ta cecha odróżnia technologię od ludzkiej psychiki.

Ostatnie dziesięciolecia zmuszały nas do błyskawicznego przystosowywania się do nowych typów interakcji socjalnych w tempie, które zapewne przyjdzie kiedyś uznać za fizjologicznie niezgodne z naszą naturą. Nadchodzi kolejna rewolucja informacyjna, przy której wszystkie poprzednie mogą się okazać łagodnym preludium. Nie wiem, czy jesteśmy w stanie przygotować się na nią, ale próbować trzeba.

Co to takiego inteligencja

Określeniem „inteligentny" posługujemy się dziś bez zająknięcia w stosunku do wielu urządzeń bądź systemów, których o autentyczną (ludzką) inteligencję nie posądzamy. Mówimy inteligentny dom, inteligentny telefon (smartfon), inteligentny samochód. Gdy pobłażliwie przypisujemy inteligencję ewidentnie bezmyślnym przedmiotom, to nie niepokoimy się, że stwarzamy sobie konkurencję sztucznych intelektualistów, albowiem wizja, że rzeczone przedmioty opanują nas wykraczając poza zakres ich żałośnie wąskiej specjalizacji jest absurdalna. Nie niepokoimy się także cokolwiek bardziej nam pokrewną inteligencją zwierząt domowych. Tutaj brak zagrożenia wynika z prostego faktu, że dystans między nimi a nami nie zmienił się od tysiącleci i nie podejrzewamy istnienia mechanizmów, które chciałyby ten stan rzeczy zaburzyć. Produkowane przez nas urządzenia stają się jednak coraz inteligentniejsze,

[6] https://openai.com/blog/chatgpt .
[7] https://www.gutenberg.org/ .

podczas gdy my zasadniczo stoimy w miejscu, a nawet, jak twierdzą niektórzy, podlegamy regresji,[8] co nawiasem mówiąc powinno nas niepokoić w stopniu nie mniejszym niż zagrożenie przez sztuczną inteligencję.

Próbując porównywać naszą ludzką inteligencję z inteligencją sztuczną musimy postawić na jednej płaszczyźnie coś co sami stworzyliśmy (i czego konstrukcję doskonale znamy) z czymś, co stworzył ktoś inny nie udostępniając nam dotąd najciekawszych schematów. Na temat identyfikacji naszego twórcy (oraz jego ewentualnych zamysłów) opowiadają wolumeny, z którymi mój skromny tekst nie poważyłby się zasiąść do wspólnego panelu. Szczęśliwie nie jest to potrzebne ani nawet użyteczne, przynajmniej przez większość naszych wywodów – nim dobrniemy do metafizycznych konkluzji, gdzie i tak zatrzymamy się przed progiem światopoglądowych spekulacji.

W naszych porównaniach musimy postępować jak inżynier. Nawet jeśli nie unikniemy spekulacji, to musimy je uprawiać od właściwego końca, skupiając się przede wszystkim na tych aspektach ludzkiej inteligencji, które (jak nam się wydaje) rozumiemy. Niezależnie od kwestii zasadniczych, istnieje dostatek aspektów niezrozumiałych i jeszcze więcej takich, które budzą zbyt wiele kontrowersji wśród specjalistów, by inżynier czuł się pewnie przy ich interpretowaniu.

Hasło „Inteligencja" rozpoczyna się w Wikipedii następującym akapitem:[9]

> „Inteligencja (od łac. intelligentia – zdolność pojmowania, rozum) – zdolność do postrzegania, analizy i adaptacji do zmian otoczenia. Zdolność rozumienia, uczenia się oraz wykorzystywania posiadanej wiedzy i umiejętności w różnych sytuacjach. Cecha umysłu warunkująca sprawność czynności poznawczych, takich jak myślenie, reagowanie, rozwiązywanie problemów."

Trudno uznać tę definicję za praktycznie użyteczną, gdyż terminy jak „pojmowanie", „rozumienie" i „umysł" nadają jej charakter cykliczny (circulus in definiendo). To co następuje jest jeszcze mniej poręczne, gdyż mowa tam o różnych rodzajach inteligencji, jak np. inteligencja językowa, wizualno-przestrzenna, muzyczna, interpersonalna czy przyrodnicza, bez ustalenia na wstępie, o co w tym wszystkim chodzi. Krótko mówiąc – całe to objaśnienie to taka sobie poezja.

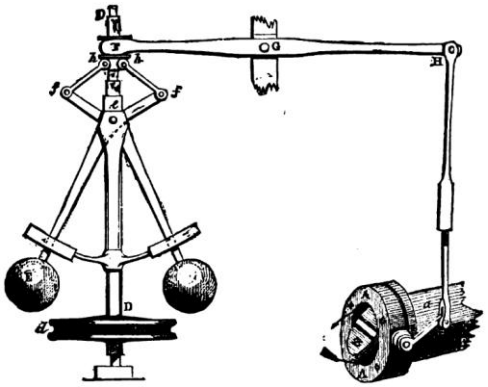
Przypisując elementy inteligencji niekontrowersyjnie bezmyślnym urządzeniom i ich konglomeratom sugerujemy się zwykle ich celową „inteligentną" reaktywnością. Przystający fragment powyższej definicji to „zdolność do postrzegania, analizy i adaptacji do zmian otoczenia". Inteligentny dom wyposażony jest w czujniki (sensory) postrzegające otoczenie (światło, temperaturę, ruch), takoż system analizy (komputer) oraz środki adaptacji (aktywatory, przełączniki). Kombinacja tych elementów realizuje pewien sensowny plan (program), którego rezultaty jesteśmy skłonni uznać za zachowanie noszące znamiona inteligencji. Takie rozumienie inteligencji przejawianej przez mechanizmy bierze się z cybernetycznych fascynacji z początków ery informacji. Największe wpływy owych fascynacji, reprezentowanych biblią mechanicznej inteligencji popełnioną przez Norberta Wienera,[10] zaznaczyły się w sferze popularyzatorskiego nagłaśniania celów rodzącej się informatyki powodując niejakie zamieszanie. Gdy jako student matematyki i informatyki rozpocząłem własną edukację w tym zakresie, czułem się rozczarowany, że nikt nie próbuje nauczać mnie cybernetyki.

Mówimy o automatycznym sterowaniu. Ambitny tytuł pracy Wienera sugerował, że osiągamy coś więcej, mianowicie zrozumienie mechanizmów leżących u podstaw zachowania wszystkiego, co się rusza, w tym oczywiście zwierząt i ludzi (człowiek, jako honorowe zwierzę, nie musiał nawet figurować w

[8] Julian Cribb, Idiocracy: is the decline in human intelligence undermining democracy? https://mahb.stanford.edu/blog/idiocracy-is-the-decline-in-human-intelligence-undermining-democracy/ .
[9] https://pl.wikipedia.org/wiki/Inteligencja .
[10] Norbert Wiener, Cybernetyka, czyli sterowanie i komunikacja w zwierzęciu i maszynie, PWN, 1971.

tytule). Może tak było, może nie. Filozofowie pozostają do dziś skłóceni – chyba nawet bardziej dziś niż w owych czasach, kiedy to każdemu wypadało się fascynować nagłym postępem.

Mechanicznym pierwowzorem systemów automatycznego sterowania był odśrodkowy regulator obrotów maszyny parowej (Rysunek 2) na którego temat James Clerk Maxwell napisał pierwszą w historii pracę naukową o teorii sterowania[11] wyprzedzając cokolwiek Wienera. W odróżnieniu od naszych bardziej praktycznych współczesnych studentów, Maxwell nie uznał za stosowne podszkolić się w marketingu. Nie potrafił nawet należycie podekscytować się swoim najważniejszym osiągnięciem,[12] bez którego nie mogłaby się rozpocząć prawdziwa elektronika i telekomunikacja.



Rysunek 2. Odśrodkowy regulator obrotów maszyny parowej (Discoveries and Inventions of the Nineteenth Century, R. Routledge, 1900).

Ustawienie archaicznego, mechanicznego regulatora pary w jednym szeregu ze współczesnymi cudami elektroniczno-informatycznej technologii inteligentnych systemów jest celowe. Zapowiada ono i uwypukla pewną wspólną cechę wszystkich takich urządzeń, którą będziemy niebawem omawiać bardziej szczegółowo. Pomimo olbrzymiej technicznej komplikacji współczesnych inteligentnych „analizatorów", ich związek z wirującym zestawem rozcapierzających się kulek opiewanym przez Maxwella jest zaskakująco bliski, podobnie jak zaskakująco bliski jest związek współczesnego komputera z liczydłem. Z pewnego zasadniczego punktu widzenia, przez te wszystkie lata szaleńczego postępu, nic istotnego się nie zmieniło, choć niby zmieniło się tak wiele.

W odniesieniu do człowieka, inteligencję oceniamy opierając się na czymś więcej niż mechanistyczna reaktywność na bodźce środowiska. W przypadku inteligentnych systemów informatycznych, włączając większość zastosowań komputerów, ich inteligentne zachowanie wymaga interpretacji ze względu na formę kontaktu z człowiekiem. Ten dystans komunikacyjny pozwalał nam przez długi czas traktować powiększające się możliwości komputerów jako proces przebiegający poza sferą naszych naturalnie ludzkich aktywności, proces, który nie mógł doprowadzić do bezpośredniej konfrontacji naszych wydolności „intelektualnych" ze względu na ewidentną różnicę w formie wyrazu oraz równie ewidentną bezsilność komputera pozbawionego naszego przewodnictwa. Przekonanie to umacniane było wczesnymi próbami konwersacji z komputerem, czyli zmuszaniem go, by opanował język naturalny,[13] co okazało się trudniejsze niż sugerowały wczesne, entuzjastyczne prognozy. Ocena inteligencji osobnika, z którym wdajemy się w interakcję rozpoczyna się od stwierdzenia, czy potrafi on budować sensowne zdania w odpowiedzi na nasze pytania. Czy potrafi zmienić temat? Czy powie nam coś ciekawego? Czy zrozumie, o co nam chodzi? Czy potrafi nas czegoś nauczyć? Nie jest przy tym absolutnie konieczne,

[11] James C. Maxwell, On Governors. Proceedings of the Royal Society of London, 16 270-283, 1868.

[12] James C. Maxwell, A Dynamical Theory of the Electromagnetic Field, Philosophical Transactions of the Royal Society of London, 155, 459-512, 1865.

[13] Przez „język naturalny" rozumieć będziemy (dowolny) język, którym na co dzień posługują się ludzie, w odróżnieniu np. od języka programowania (które to pojęcie potocznie kojarzy się z komputerami).

byśmy komunikowali się głosem. Wymiana notatek przez email, WhatsApp, SMS w zupełności wystarczy. Wystarczą także próbki pisemnej lub artystycznej twórczości delikwenta, jeżeli jesteśmy w stanie stwierdzić ich autentyczność. Wczesne próby przymuszenia komputerów do inteligentnych konwersacji wypadały raczej kiepsko – aż do dziś.

Konwersacje z maszyną i test Turinga

Produktywna interakcja człowieka z komputerem rzadko przypomina rozmowę dwojga ludzi, pomimo dostępności w dzisiejszych czasach zaawansowanych narzędzi rozpoznawania mowy i syntezy dźwięku. Jako technologiczny dinozaur, nie jestem entuzjastą wykrzykiwania poleceń mojemu komputerowi, czy nawet telefonowi, choć w tym drugim przypadku zdarza mi się chodzić na ustępstwa. Tak więc klepię teraz te zdania w procesor tekstu posługując się klawiaturą, choć teoretycznie mógłbym je dyktować. Pewnie nawet jakoś by szło. Niektórzy tak pracują. Nawet ja, gdy podczas spaceru z psem coś mi przypadkiem przyjdzie na myśl, potrafię podyktować do telefonu notatkę, którą aplikacja zamieni na (zwykle pokrętny i z trudem zrozumiały) tekst.

Programy, których najczęściej używamy (przynajmniej do poważnych celów) posiadają interfejs o charakterze imperatywnym: użytkownik zleca, komputer realizuje, gdzie repertuar możliwych akcji użytkownika określony jest przez zestaw prostych komend, pól do wypełnienia tekstem, miejsc do kliknięcia myszą. Nawet jeśli współczesne aplikacje próbują czasem zgadnąć, o co nam chodzi, często wolimy na tym nie polegać, szczególnie jeśli zależy nam na poprawnym wyniku. „Inteligentna" klawiatura w moim smartfonie płata czasem zabawne figle próbując zgadnąć słowo, które wyślizguję po niej palcem, co powoduje, że redagowaną notatkę sprawdzam kilkakrotnie przed wysłaniem jej poważnemu odbiorcy.

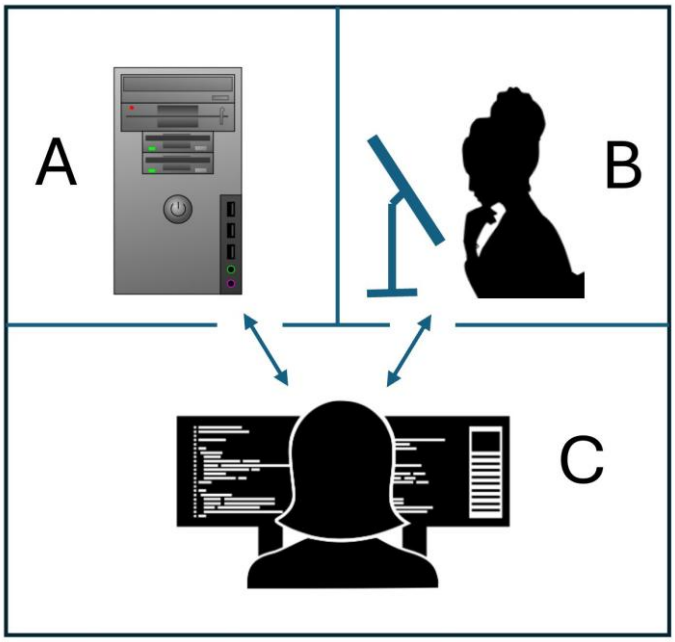
Od samego początku istnienia informatyki, zanim dyscyplina uzyskała nazwę, jej wizjonerzy wyobrażali sobie przyszłe konwersacje z komputerem w języku naturalnym. Wydawało się oczywiste, że odpowiednio pojemny komputer wyposażony w odpowiednio duży i wymyślny program, będzie potrafił rozpoznawać zdania i odpowiadać na nie sensownie. Może nie od razu i nie na wszystkie, ale nietrudno było zacząć. Jeden z historycznie pierwszych ważnych języków programowania, COBOL,[14] posiadał składnię przypominającą proste zdania w języku naturalnym. Rzecz jasna, odstępstwa (pomyłki) od sztywnej składni nie były tolerowane, pomimo oczywistości intencji człowieka programisty. Tym niemniej, ta wczesna radość z możliwości zlecania komputerowi zadań w formie przypominającej zwykły ludzki język przypominała satysfakcję, jakiej doznajemy wyuczywszy psa nowej sztuczki.

Już w roku 1950, Alan Turing, jeden z twórców teoretycznej informatyki, zaproponował test, mający (według potocznej interpretacji) ustalić, czy komputer myśli.[15] Powodem, dla którego idea takiego testu pojawiła się w polu zainteresowań nauki stało się teoretyczne sformułowanie i zrozumienie ogólnego pojęcia obliczenia, czyli klasy zadań potencjalnie realizowalnych przez komputer działający według zasad von Neumanna. Zainspirowało to spekulacje na temat granic możliwości takich urządzeń, a ściślej mówiąc, możliwości wykonywanych przez nie programów. Pomimo tragicznie skromnej wydolności ówczesnego sprzętu, wizjonerzy nie czuli się ograniczeni stanem technologii próbując przewidzieć, dokąd to wszystko może nas kiedyś doprowadzić.

Na temat testu Turinga wypisano mnóstwo popularnych komentarzy oraz interpretacji, z których (delikatnie mówiąc) nie wszystkie uczciwie oddają literę pomysłu. Turingowi nie chodziło o to, czy komputer faktycznie jest w stanie myśleć, w ludzkim sensie, cokolwiek by to miało znaczyć. Nie próbował on dotykać filozoficznych i metafizycznych aspektów ludzkiego myślenia, świadomości, inteligencji, zdając sobie sprawę, że nie posiadł klucza do formalizacji tych pojęć. Jego proponowany eksperyment

[14] Mo Budlong, COBOL, Wydawnictwo Helion, 2002.
[15] Alan Turing, Computing Machinery and Intelligence, Mind, LIX (256), 433-460, 1950.

przedstawiał się tak. Mamy dwa zamknięte pomieszczenia; w jednym znajduje się komputer a w drugim człowiek. Pomieszczenia połączone są z zewnętrznym światem kanałami (na przykład kablami) pozwalającymi przesyłać, w obie strony, pisaną informację. Możemy sobie wyobrazić, że każdy kanał kończy się czymś w rodzaju monitora z klawiaturą, co w tamtych czasach nazywało się dalekopisem. Przy monitorach siedzi człowiek-tester i komunikuje się z lokatorami pomieszczeń prowadząc z nimi dowolną konwersację – według własnego wyboru. Celem testera jest odgadnięcie, w którym pomieszczeniu znajduje się komputer. Jeśli statystycznie, w długiej serii prób, sukces testera okaże się porównywalny do wyniku rzutu monetą, wówczas przyjdzie uznać, że pod względem zdolności konwersacyjnych komputer jest nieodróżnialny od człowieka. Wypadnie zatem pragmatycznie przypisać mu ludzką inteligencję.



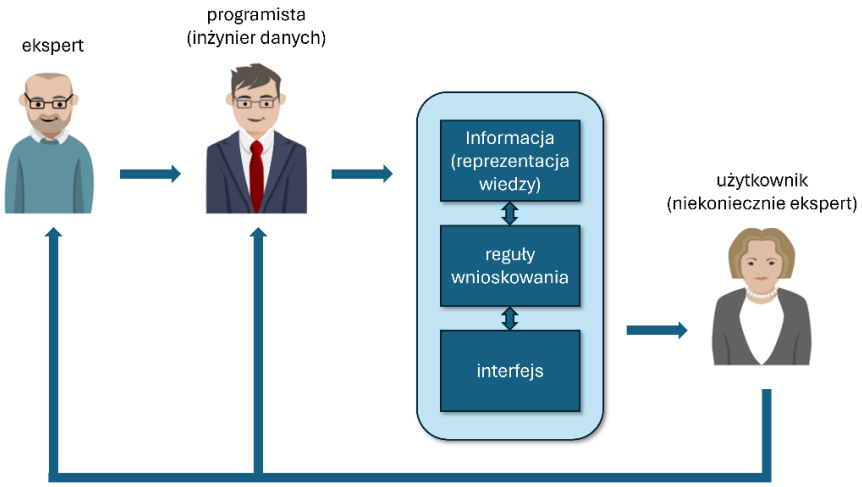
Rysunek 3. Test Turinga.

Jednym zdaniem, chodzi o to, by w konwersacji z człowiekiem komputer nie dawał się łatwo odróżnić od człowieka. Zauważmy, że na użytek testu Turinga nie interesuje nas precyzyjny zakres specyficznych możliwości komputera, który z samej natury (i od samego początku) daje mu przewagę nad człowiekiem. Na przykład, rozwiązując zadania arytmetyczne komputer z łatwością wykaże się sprawnością (nieomylnością) przerastającą możliwości ludzkie, co oczywiście mogłoby posłużyć do jego zdemaskowania. Przy pragmatycznym podejściu do testowania nie musi to być istotne. Tester może się uczciwie powstrzymać od egzaminowania rozmówców z arytmetyki. Program komputera może też celowo zwodzić testera względem trudności sprawianych mu przez obliczenia czy ilości czasu potrzebnego mu na wyprodukowanie odpowiedzi. Nie chodzi nam o pozbawienie komputera przyrodzonych mu cech, lecz o ustalenie, czy potrafi on skutecznie emulować ludzką inteligencję. Test Turinga skupia się na najistotniejszym praktycznym elemencie tej emulacji, jakim jest inteligentna konwersacja komputera z człowiekiem z pominięciem technicznej strony wymiany informacji sprowadzonej w formule testu do pisania znaków na klawiaturze i odczytywania napisów na monitorze. Abstrahuje on w ten sposób od wszelkich dekoracyjnych aspektów interakcji z komputerem nieistotnych z punktu widzenia „czystej” inteligencji, jak na przykład opcja rozpoznawania mowy i posługiwania się głosem. Algorytm realizujący sztuczną inteligencję komputera, w swojej esencji, operuje na zdaniach języka naturalnego. Jego celem jest rozpoznanie (zrozumienie) zdań nadchodzących od testera, osadzenie ich w kontekście prowadzonej rozmowy i wygenerowanie własnych zdań stanowiących jej dorzeczną kontynuację.

W czasach, gdy test Turinga był najintensywniej dyskutowany i powodował największe emocje, nie istniały przesłanki technologiczne dla prób jego zaliczenia przez komputer. Tym niemniej, w niektórych sferach naukowych związanych ze sztuczną inteligencją wytyczał on od samego początku wektor postępu. Tworzono programy zdolne do ograniczonych konwersacji w języku naturalnym, np. osławiony system Eliza,[16] inspirujący wielu badaczy na całym świecie. Pomimo poważnych wysiłków akademickiej społeczności i mnóstwa usprawnień żmudnie wprowadzanych do tych programów przez lata nie zbliżały się one w żadnym istotnym stopniu do upragnionego celu, nawet przy drastycznym zawężaniu tematów konwersacji.

Przypomina mi się historia z zamierzchłego okresu moich studiów, gdy w roku 1973 nasza wydziałowa grupa sztucznej inteligencji demonstrowała swój własny system konwersacyjny o nazwie Marysia.[17] Marysia emulowała sprzedawczynię w sklepie; można ją było zapytać o towar i przeprowadzić prostą rozmowę związaną z zakupem. Jeden z moich kolegów podszedł do klawiatury i napisał: „Dzień dobry, czy jest szynka?". „Szynki nie ma" – padła zwięzła (niezbyt zaskakująca w owych czasach) odpowiedź. Kolega nie dał za wygraną: „Czy jest prawdą, że nie ma szynki?". „Prawdy też nie ma" – poinformowała Marysia.

Poczynając od połowy lat 70-tych, poprzez lata 80-te i 90-te, ciężar badań w sztucznej inteligencji nakierowanych na usprawnianie konwersacji człowieka z maszyną przenosił się w stronę tzw. systemów ekspertowych,[18] gdzie odstąpiono od abstrakcyjnego celu wytyczonego przez test Turinga skupiając się na praktycznych zastosowaniach. Dążenie do iluzji obcowania z człowiekiem zeszło na drugi plan, a zakres komputerowej inteligencji został zredukowany do wąskiej specjalistycznej ekspertyzy. Historycznym przykładem jest tu system MYCIN,[19] stworzony na uniwersytecie w Stanford w 1972, służący do diagnozowania infekcji bakteryjnych. Stanowił on pierwowzór wielu medycznych systemów diagnostycznych oraz innych systemów ekspertowych tworzonych w kolejnych latach.



Rysunek 4. Struktura systemu ekspertowego.

Konwersując z systemem ekspertowym, użytkownik nie był zainteresowany złudzeniem interakcji z człowiekiem i w szczególności nie oczekiwał od komputera ludzkiej poczciwości, niepewności, czy okazjonalnych pomyłek. System funkcjonował w oparciu o reguły oraz wpojoną mu bazę wiedzy, którą

[16] Joseph Weizenbaum, ELIZA – a Computer Program for the Study of Natural Language Communication between Man and Machine, Communications of the ACM, 9, 36-35, 1966.
[17] Janusz S. Bień, Założenia Polskiego Systemu Konwersacyjnego Marysia, Wydawnictwa UW, 1973.
[18] Mark Stefik, et al., The Organization of Expert Systems, a Tutorial, Artificial Intelligence, 18 (2), 135-173, 1982.
[19] Edward H. Shortliffe, Computer-Based Medical Consultations: MYCIN, Artificial Intelligence Series, 2, Elsevier, 1976.

można było uzupełniać (korygować) na podstawie doświadczenia. Tryb interakcji był zwykle wymuszony przez system. Na przykład, w przypadku systemu diagnostycznego, komputer pytał a użytkownik odpowiadał, co przypominało rozmowę z lekarzem. W ramifikacjach ekspertyzy systemu, interakcja taka mogła być uważana za inteligentną z praktycznego, pragmatycznego punktu widzenia.

Komputerowe tłumaczenie

Historycznie, najbardziej pociągającym zastosowaniem komputerów związanym z przetwarzaniem języka naturalnego było tłumaczenie tekstów z jednego języka na drugi. Na pierwszy rzut oka wydaje się, że problem jest dobrze określony i że proces tłumaczenia można próbować opisywać przy pomocy reguł zrozumiałych przez komputer. Słownik oraz podręcznik gramatyki to dokumenty, które (przynajmniej w sferach humanistów) uchodzą za precyzyjne. Gdy jako sześcioletnie dziecko rozpoczynałem naukę języka angielskiego, fascynowała mnie zabawa polegająca na zastępowaniu w zdaniach polskich słów ich angielskimi odpowiednikami. Wydawało mi się, że produkuję w ten sposób całkowicie legalne angielskojęzyczne odpowiedniki polskich zdań. Komputer zaopatrzony w słownik, zestaw reguł gramatycznych oraz stosunkowo prosty program potrafi tyle samo. Niestety, szybko wyda się, że jest to zabawa dziecinna i całkowicie pozbawiona użyteczności. Pośród kanadyjskiej Polonii w miejscu mojej emigracji popularne były dowcipne rodzynki słownikowego tłumaczenia w rodzaju: „thank you from the mountain", czy „roads father, I have long". Każdy, kto kiedykolwiek próbował nauczyć się obcego języka rozumie doskonale, że gdyby słownik wzbogacony o garstkę mechanicznych reguł stanowił wystarczający ekwipunek tłumacza, wszyscy bylibyśmy poliglotami.

Rysunek 5. Elon Musk wnoszący zlew do biura Twittera.

Najtrudniejszym elementem wyposażenia programu, który tłumaczyłby skutecznie jest wiedza na temat niesystematycznych własności języka pozwalająca wykrywać gramatyczne wyjątki, idiomy oraz subtelne różnice w znaczeniu wynikające z kontekstu, który czasem może być bardzo długi i trudny do automatycznego wychwycenia. Gdy parę miesięcy temu przygotowywałem prezentację na temat historii komputerowego tłumaczenia, umieściłem na tytułowym slajdzie zdjęcie Elona Muska, wycięte z klipu wideo, który kilka tygodni wcześniej, kiedy to Musk sfinalizował zakup Twittera, obiegł Internet (patrz Rysunek 5).[20] Na fotografii nowy właściciel wkracza do budynku zarządu firmy dźwigając kuchenny zlew. Opis obrazka brzmi: „Let that sink in". To krótkie zdanie, jeśli pozbawimy je kontekstu,

[20] https://twitter.com/elonmusk/status/1585341984679469056 .

może znaczyć „Wpuście ten zlew” lub „Niech to do was dotrze”. Obie opcje wynikają wprost ze słownikowej interpretacji słów oraz reguł gramatyki.

Zastanówmy się jakiego rodzaju wyzwanie oczekuje tłumacza (wszystko jedno człowieka czy maszynę), którego zadaniem jest przetłumaczenie podpisu pod fotografią. Dopuszczając dwie drastycznie różne możliwe interpretacje prostego zdania, wypada przyjrzeć się zdjęciu w poszukiwaniu kontekstu. Myśląc o komputerze, przyjdzie nam zatem od razu zauważyć, że powinniśmy umożliwić mu interpretowanie zawartości obrazków, co bez wątpienia powiększa rozmiar problemu. Zakładając, że rozpoznaliśmy tam zlew, pierwszy wariant tłumaczenia nasuwa się natychmiast. Nawiasem mówiąc, gdyby dwadzieścia lat temu udało nam się stworzyć program, który potrafiłby samodzielnie dojść do takiej konkluzji, moglibyśmy liczyć na uznanie kolegów i spory huk w akademickiej prasie. No ale to przecież nie wszystko. Jeśli przypadkiem zauważyliśmy, że osobą wnoszącą zlew jest Elon Musk, a w hallu budynku widnieje logo Twittera, to mamy podstawy podejrzewać, że chodzi tu o coś więcej niż pozwolenie na instalację nowego zlewu. Aby jednak zrozumieć problem do końca, musimy najzwyczajniej w świecie wiedzieć co jest grane, czyli znać intrygę i jej polityczne konotacje. Jeśli przypadkiem nic jeszcze nie wiemy (co należałoby założyć w sytuacji komputera), to być może dodatkowy kontekst da się pozyskać z obszerniejszego tekstu komentującego klip. Przestudiowawszy go dojdziemy do wniosku, że tak naprawdę chodzi o znaczenie drugie, a ów nieszczęsny zlew służy jedynie za dowcipny rekwizyt. No i jako profesjonalny tłumacz pragnący przekazać czytelnikowi jak najwięcej ze znaczenia tłumaczonej frazy w jej pełnym kontekście, zmuszeni będziemy opowiedzieć mu o symbolicznej roli zlewu, która w języku docelowym (np. polskim) nie jest bezpośrednio zrozumiała.

Na opisanym przykładzie widzimy jak próba sensownego przetłumaczenia krótkiego komunikatu, gdzie przychodzi zdecydować między dwoma prostymi wariantami tłumaczenia, wymaga wnikliwego zrozumienia kontekstu. Intuicyjnie, umiejętność rozstrzygania podobnych kwestii wydaje się być warunkiem koniecznym dla emulowania ludzkiej inteligencji na poziomie pozwalającym maszynie istotnie zbliżyć się do zaliczenia testu Turinga. Zauważmy, że tłumaczenie zostało w przykładzie zredukowane do roli pretekstu, gdyż surowy problem tłumaczenia jest zasadniczo słownikowy; identyfikacja dwóch dopuszczalnych wersji wynikowego tekstu nie sprawia kłopotów. Mówiąc inaczej, warunkiem koniecznym dobrego tłumaczenia na poziomie ludzkiego eksperta jest wstępne opanowanie „rozumienia” kontekstu. Problem jest niezależny od tłumaczenia i jego rozwiązanie niewątpliwie ułatwi komputerowi produkowanie dorzecznych tekstów na różne okazje, niekoniecznie związane z tłumaczeniem. Nie chodzi nam o to, by komputer (program) potrafił rozumieć i interpretować świat w ludzkich kategoriach (czego sformalizować nie potrafimy), lecz by przejawiał satysfakcjonujące aspekty takiego zachowania przy posługiwaniu się językiem naturalnym.

Powyższy przykład ma prawo deprymować, gdyż sugeruje, że bez pozyskania semantycznej wiedzy o świecie, komputer nie potrafi wypowiadać się o nim w sposób sensowny. Nawet jeśli przyuczymy go do produkowania w miarę poprawnych zdań w języku naturalnym, to co najwyżej uczynimy z niego coś w rodzaju papugi; nie będzie on w stanie korzystać z tej umiejętności dla poprawnego przekazywania treści.

To co dla jednych było przez długi czas deprymujące, dla innych stanowiło kojące zapewnienie permanentnej niższości komputera nad człowiekiem. Początkowy entuzjazm prób mechanicznego tłumaczenia tekstów naukowych z języka rosyjskiego na angielski i odwrotnie zainicjowanych na początku lat 60-tych (przy niejakim dopingu natury politycznej) doprowadził do głębokiego rozczarowania odzwierciedlonego w słynnym raporcie komisji ALPAC[21] rekomendującym zaniechanie kosztownych wysiłków

[21] John R. Pierce, et al. Language and Machines — Computers in Translation and Linguistics. ALPAC report, National Academy of Sciences, National Research Council, Washington, DC, 1966.

w tym zakresie jako nieopłacalnych i pozbawionych praktycznych perspektyw. Na dłuższą metę okazało się jednak, że deprymujący charakter podchwytliwych problemów lingwistyczno-semantycznych jest przejściowy. Obecne systemy mechanicznego tłumaczenia, jak DeepL,[22] produkują teksty o jakości porównywalnej z płodami zawodowych tłumaczy, choć jeszcze nie tych najlepszych, powiedzmy literackich.

Pozostańmy jeszcze przez chwilę przy komputerowym tłumaczeniu, gdyż jest to dziedzina dobrze ilustrująca genezę i zasady współczesnych technik produkowania inteligentnych tekstów przez komputery. Bardzo wcześnie zrozumiano, że droga do sukcesu nie wiedzie przez próby wyposażenia komputera w słownik oraz zrozumiałą przezeń reprezentację gramatycznych i heurystycznych reguł. Cóż zatem pozostaje? Skoro wszystkie istotne problemy mechanicznej translacji biorą się z niestosowalności ogólnych reguł do złośliwych przypadków szczególnych, potencjalnie obiecującym rozwiązaniem byłoby oparcie się na wyczerpującym zestawie przykładów poprawnych tłumaczeń wykonanych przez ludzkich tłumaczy-ekspertów, możliwie obejmujących jak największy zakres spisanej wiedzy.

Zauważmy, że dziecko ucząc się swojego pierwszego języka nie zaczyna od opanowania słownika i przez długi czas nie rozumie, co to gramatyka. Na samiutkim początku swojej edukacji nie zna ono żadnego słowa i żadnej reguły będąc otwarte na absolutnie dowolny z ludzkich języków. Uczy się zasadniczo przez obserwację i własne próby lingwistycznych interakcji z otoczeniem, czyli przez przykłady.

Nie chciałbym tu nadużywać popularnej analogii między sposobem nabywania językowej biegłości przez dziecko i przez współczesny komputerowy model języka, gdyż istnieją poważne różnice, które uwypuklimy później. Należy zdawać sobie z nich sprawę od samego początku, by uniknąć odruchu antropomorfizacji komputerów przejawiających powierzchownie ludzkie cechy. Celem naszej analogii jest jedynie obserwacja, że nauka języka przez przykłady posiada pewien naturalny sens. Sposób, w jaki ludzie opanowują pierwszy język jest konsekwencją (i jednocześnie przesłanką) jego struktury, gdzie w pewnym fundamentalnie ważnym sensie (z punktu widzenia skutecznego opanowania języka) rządzą wyjątki, które niestety, w niezgodzie z obiegowym powiedzeniem, nie potwierdzają reguł, a przynajmniej nie dają się z nich mechanicznie wydedukować. Powoduje to, że jedyny rozsądny sposób, w jaki komputer może próbować opanować język naturalny prowadzi w tym samym kierunku co w przypadku dziecka, choć cokolwiek odmienną drogą. Droga ta jest odmienna nie tylko dlatego, że tak naprawdę to nie rozumiemy, w jaki sposób my uczymy się (pierwszego) języka, nie mówiąc już o braku formalizmów w tym zakresie, które dałoby się przełożyć na komputerowe algorytmy. Doskonale natomiast rozumiemy, że komputer nie może nas żywcem emulować w procesie nabywania kompetencji językowych, pomimo (a także na skutek) olbrzymiej przewagi, jaką nad nami posiada: elektronicznej pamięci, która nie zapomina oraz procesorów pozwalających mu wykonywać bezbłędnie miliardy operacji na sekundę. Autorytety od lingwistyki, na przykład Noam Chomsky,[23] twierdzą, że umiejętność szybkiego i automatycznego uczenia się języka przez dziecko jest jedną z przyrodzonych cech rodzaju ludzkiego, coś co w szczególności umożliwiło powstanie cywilizacji. Dziecka nie trzeba uczyć mowy podobnie jak nie trzeba je uczyć chodzenia.

Teoretycy próbujący powiązać swoje abstrakcyjne wizje z rzeczywistością uciekają się czasem do formalnych spekulacji zwanych eksperymentami myślowymi.[24] Celem takiego eksperymentu jest wykazanie, że jakiś proces lub zjawisko jest dopuszczalne z punktu widzenia teorii, choć jego praktyczna

[22] https://www.deepl.com/pl/translator .

[23] Noam Chomsky, Syntactic Structures, De Gruyter Mouton, 2009.

[24] W nauce rozważanie tego typu znane jest pod nazwą Gedankenexperiment, które to słowo używane jest powszechnie w językach innych niż niemiecki, np. w angielskim, prawdopodobnie dla oddania honoru najwybitniejszym niemieckim fizykom teoretycznym, którzy z upodobaniem (i sporymi sukcesami) unikali brudzenia sobie rąk rzeczywistymi eksperymentami pozostawiając je majsterkowiczom (i z góry znając ich wyniki).

realizacja mogłaby się rozbić o przyziemne problemy niegodne uwagi teoretyka, jak brak pieniędzy, czasu, personelu czy innych zasobów. Dopuszczalność wynika z przepisu, gdzie wszystkie składniki są poprawne (logicznie i fizycznie), wszystkie akcje wykonalne z zasady, choć całość może być niemożliwa do praktycznego zrealizowania ze względów pozaformalnych. Albert Einstein, gdy w wieku 16 lat przymierzał się powoli do sformułowania szczególnej teorii względności, rozważał niepokojący go eksperyment myślowy, w którym on (znaczy się Albert Einstein) dogania falę elektromagnetyczną, czyli na przykład promień światła, i zrównawszy się z nim konstatuje, że pola elektryczne i magnetyczne przestają oscylować w jego ramce odniesienia.[25] Jasne, że Einstein nie mógł pieszo dogonić fali elektromagnetycznej, nawet jako nastolatek, lecz według ówczesnych teorii nie było formalnych przeszkód, by dokonać tego posługując się wyjątkowo szybkim pojazdem (który oczywiście także był technicznie niewykonalny). Szczęśliwie nie było potrzeby inwestowania w przemysł rakietowy, by sprawdzić (i rozwiać lub potwierdzić) niepokoje młodego Einsteina, gdyż wynik eksperymentu myślowego stanowił logiczną konsekwencję obowiązujących wówczas teorii wykazując drogą dedukcji, że coś jest z nimi źle.

Załóżmy dla ustalenia uwagi, że chodzi nam o tłumaczenie z angielskiego na polski i wyobraźmy sobie, w ramach eksperymentu myślowego, że wszystkie możliwe teksty zostały już przetłumaczone (przez ludzkich tłumaczy-ekspertów) i że w olbrzymiej bazie danych przechowujemy wszystkie możliwe oryginały razem z ich tłumaczeniami. Aby spełnić warunek dopuszczalności, ograniczymy się do tekstów o długości powiedzmy miliona znaków. Pokryje to praktycznie wszystkie teksty, jakie ktoś może kiedykolwiek chcieć przetłumaczyć i zagwarantuje, że ich liczba (choć niewyobrażalnie wielka) jest formalnie skończona. Wysiłek ludzkich tłumaczy potrzebny dla utworzenia naszej bazy danych jest oczywiście poza wszelką szansą na praktyczną realizację, ale jedynym banalnym i nieistotnym problemem jest niedostatek kadry (co można porównać z nieważnym utrudnieniem, jakim dla młodego Einsteina było osiągnięcie prędkości światła). Komputer posiada dostęp do naszej monstrualnej bazy danych i otrzymawszy tekst do przetłumaczenia, wyszukuje oryginał, po czym produkuje odpowiadające mu tłumaczenie. Proces taki przypomina wyszukiwanie słów w olbrzymim słowniku, gdzie słowami są pełne teksty źródłowe oraz ich gotowe translacje.

Jakkolwiek niedorzecznie przedstawiałby się projekt programu tłumaczącego skonstruowanego według powyższego pomysłu, rzeczywiste współczesne systemy tłumaczenia można uznać za sprytne próby aproksymowania jego idei w oparciu o realistyczne zasoby. Początkowym założeniem jest dostępność rozległego zestawu przykładów, czyli tak zwanego korpusu tekstów. W epoce rozkwitu Internetu, stanowiącego olbrzymie repozytorium informacji wygenerowanej przez naszą cywilizację, wiadomo, dokąd się udać by go pozyskać, być może z niejakim trudem polegającym głównie na odsiewaniu ziarna od plew. Interesują nas dwa typy korpusów: monojęzyczny obejmujący przykłady poprawnych tekstów w jednym określonym języku oraz dwujęzyczny (zwany też równoległym) zawierający przykłady tłumaczeń. Zakładając, że naszym celem jest tłumaczenie z angielskiego na polski, korpus monojęzyczny dotyczyć będzie języka polskiego (czyli docelowego), zaś korpus dwujęzyczny zawierać będzie przykłady tłumaczeń z języka angielskiego na polski. Zakres korpusu monojęzycznego, jaki potencjalnie da się wydobyć z Internetu obejmuje w dzisiejszych czasach praktycznie wszystkie istotne, zarówno sensowne jak i bezsensowne teksty wypisane w danym języku, które z jakichś powodów przedostały się do sieci. W szczególności znajdują się tam zasadniczo wszystkie książki i oficjalne publikacje kiedykolwiek wygenerowane przez ludzi. Popularnym repozytorium tekstów, często na niegłupie tematy, jest Wikipedia, której całą zawartość każdy może pobrać na życzenie.[26] W zakresie języka angielskiego (który siłą rzeczy jest lepiej reprezentowany niż język polski), popularnym i rekomendowanym

[25] Albert Einstein, Autobiographical Notes. In Paul A. Schilpp, Albert Einstein; Philosopher, Scientist (2nd ed.), New York Tudor Publishing, pp. 2-95, 1951.
[26] https://en.wikipedia.org/wiki/Wikipedia:Database_download .

elementem korpusu jest wyselekcjonowany zestaw publicznie dostępnych książek zawierający starannie dobraną mieszankę gatunków.[27]

Korpus drugiego rodzaju (czyli dwujęzyczny) może być trudniejszy do pozyskania, lecz sytuacja jest lepsza niż mogłoby się wydawać na pierwszy rzut oka. Istnieją oczywiście tłumaczenia literackie (książki, sztuki teatralne, eseje), których stosowalność jako praktycznych wzorców tłumaczenia bywa ograniczona – właśnie ze względu na ich literacki charakter powodujący, że w lokalnych aspektach (potencjalnie przydatnych w sytuacjach praktycznych) tłumaczenie ma prawo znacznie odbiegać od oryginału. Z pomocą przychodzi tu biurokracja Unii Europejskiej dostarczając niechcący użytecznego korpusu wielojęzycznego stanowiącego ulubioną pożywkę badaczy i projektantów systemów automatycznego tłumaczenia.[28] Według unijnych zasad, wszystkie wystąpienia w europejskim parlamencie muszą zostać udostępnione światu jako pisane dokumenty we wszystkich językach krajów członkowskich.[29]

Ktokolwiek miałby ochotę sprawnie tłumaczyć z języka B na język A musi opanować oba języki. Zauważmy jednak, że oczekiwanym wynikiem procesu tłumaczenia jest możliwie wysokiej jakości tekst w języku A, który z dużą wiernością oddaje znaczenie oryginału. Tak więc biegłość w języku A jest ważniejsza. Z punktu widzenia języka B istotne jest jedynie solidne zrozumienie treści, czyli coś co nazwalibyśmy pasywną znajomością języka. Nie wymaga się od tłumacza by potrafił tworzyć piękne (czy nawet poprawne) zdania w języku B, gdyż formalnie nie wchodzi to w zakres oczekiwanych od niego kompetencji. Znów należy zachować ostrożność z analogią między komputerem a człowiekiem, gdyż problem wygląda nieco inaczej od strony komputera. W przypadku tłumaczenia przez człowieka, da się obronić argument, że naprawdę dobry tłumacz powinien doskonale władać oboma językami, co pozwoli mu wczuwać się we wszelkie lingwistyczne i kulturowe niuanse oryginału, ale wystarczy mu doskonale posiąść praktyczną, kulturalnie bogatą znajomość języka źródłowego nie nabywszy dostatecznych (i zbędnych dla tłumacza) kompetencji do uprawiania w nim literatury. Dla komputera, algorytmiczna realizacja procesu tłumaczenia sprawia, że rodzaj „znajomości" języka B, jaki jest mu potrzebny dla wyszukiwania odpowiednich zdań języka A jest w większości (technicznych) aspektów odmienny od wymaganego trybu działania w zakresie języka A. Mówiąc prościej, komputer nie „przeżywa" znaczenia tłumaczonego tekstu w ten sam sposób co ludzki tłumacz.

Modelowanie języka

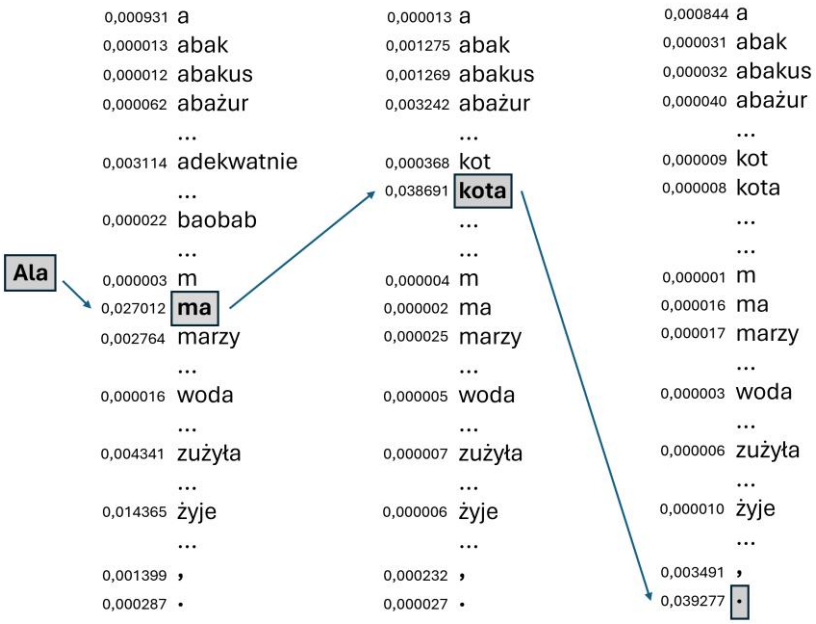
Biegłość w posługiwaniu się językiem naturalnym w celu przekazania treści rozpoczyna się od umiejętności tworzenia poprawnych zdań. To oczywiste, że zanim ustalimy, na jaki temat chcemy się wypowiedzieć w danym języku, musimy wpierw ustalić, czy w ogóle potrafimy się w nim wysławiać. Z punktu widzenia komputera, zdolność generowania poprawnych zdań w języku naturalnym stanowi warunek wstępny dla wszelkich aplikacji wymagających produkowania takich zdań w konkretnych celach. Ich przykładami są tłumaczenie tekstów oraz inteligentna konwersacja z człowiekiem.

Problem tłumaczenia z języka B na język A zawiera zatem podproblem generowania poprawnych zdań w języku A. Drugi problem, czyli problem przypisywania tekstom wejściowym w języku B tłumaczeń, sprowadzi się do zdefiniowania funkcjonalnego kryterium ograniczającego repertuar zdań języka A do takich, które wiążą się z tekstem wejściowym. Tak się szczęśliwie składa, że rozwiązawszy pierwszy problem uzyskamy coś więcej niż częściowe rozwiązanie problemu tłumaczenia, mianowicie system produkowania poprawnych zdań w języku A, który być może da się zastosować do innych celów.

[27] https://en.wikipedia.org/wiki/BookCorpus .
[28] https://www.statmt.org/europarl/ .
[29] Można mieć zastrzeżenia co do charakteru językowych i translatorskich kompetencji nabytych w oparciu o taki korpus. Nie bądźmy zatem zaskoczeni błyskotliwymi zastosowaniami sztucznej inteligencji w polityce, które nas nieuchronnie czekają.

Odpowiednio dobierając kryteria zawężające dla naszego generatora, potrafimy być może zaprząc go do inteligentnych konwersacji z człowiekiem w języku A, gdzie oczywiście napotkamy ten sam podproblem generowania ogólnie poprawnych zdań.

W taki sposób, problem komputerowego tłumaczenia, przez swoją oczywistą praktyczną użyteczność (czyli bezpośrednią wartość komercyjną) napędzał postęp w tej części sztucznej inteligencji, która miała na celu doprowadzenie do inteligentnych konwersacji w języku naturalnym między człowiekiem i maszyną. Po prostu: aby dobrze tłumaczyć należy się dobrze wysławiać. O ile to drugie można było przez pewien czas traktować jako naukową fanaberię, o tyle maszynowe tłumaczenie zawsze miało namacalny (komercyjny) sens stanowiąc ważny przedmiot zainteresowania praktycznego użytkownika.



Rysunek 6. Generatywne funkcjonowanie modelu języka. Zwróćmy uwagę, że wartości przypisane kolejnym słowom generowanej sekwencji zależą od jej dotychczasowej postaci.

System umożliwiający tworzenie poprawnych zdań w określonym języku A nazywa się modelem tego języka. Posiada on niezwykle prostą definicję formalną, którą zrozumieć potrafi każdy, niekoniecznie matematyk. Na wejściu, model otrzymuje sekwencję słów języka A, przy czym znaki interpunkcyjne traktowane są jak honorowe słowa.[30] Na wyjściu model produkuje pojedynczą liczbę, którą można interpretować jako miarę jakości sekwencji, a mówiąc odrobinę bardziej formalnie prawdopodobieństwo, że taka sekwencja pojawia się w języku. Model pozwala zatem wartościować sekwencje języka i porównywać je pod względem jakości. Na przykład, zestawiając ze sobą dwie wersje zdania o tej samej treści, używające innego szyku lub nieco innych słów, model powie nam, która wersja jest lepsza.

Taka definicja modelu abstrahuje od jego dynamiki (czy w ogóle implementacji) i skupia się na minimalnej funkcjonalności, która wystarczy dla ścisłego sformułowania wielu lingwistycznych zadań. Wyobraźmy sobie, że naszym celem jest generowanie zdań na określony temat. Ktoś rzuca słowo „Ala”. Nasz algorytm generowania zdań patrzy teraz w słownik (patrz Rysunek 6), wybiera z niego po kolei wszystkie słowa, ustawia każde z nich na końcu tak zainicjowanej sekwencji (czyli po słowie „Ala”) i pyta model o wartościowanie. Słowo, dla którego model zwróci największą wartość zatrzymamy jako

[30] Ściślej mówiąc, nie są to słowa (w potocznym sensie), lecz tak zwane „żetony” (ang. tokens). Jak zauważyliśmy, zbiór żetonów obejmuje na przykład znaki interpunkcyjne. Często bywa tak, że zlepek słów, skrót, ciąg znaków specjalnych, itd. należy traktować jako jednostkę z punktu widzenia języka. Nie będziemy komplikować naszej dyskusji przez niepotrzebne gmatwanie terminologii.

następne. Powiedzmy, że jest nim „ma”. Uzyskaliśmy w ten sposób rozszerzoną sekwencję „Ala ma”, dla której znów poszukujemy najlepszej kontynuacji, i tak dalej. Najprawdopodobniej, model wyprodukuje kolejno „kota” a potem kropkę. Tak powinien się zachować dobry model naszego ojczystego języka.

Opisana powyżej procedura jest oczywiście niezbyt efektywna i nikt przy zdrowych zmysłach nie zaimplementuje jej w taki sposób. Zauważmy przy okazji, że można ją uważać za usprawnienie pewnej wizji dostarczonej nam przez eksperyment myślowy z poprzedniego rozdziału. Zmodyfikujmy naszą monstrualną bazę danych z tamtego eksperymentu wyrzucając z niej język B i pozostawiając jedynie zestaw wszystkich tekstów (docelowego) języka A, który wynajęta armia ekspertów językowych przytnie, by zawierał tylko wzorcowe, poprawne przykłady zdań. Obliczając ranking wejściowej sekwencji, algorytm sprawdzi, ile razy pojawia się ona jako fragment tekstu obecnego w naszej bazie danych ustalając w ten sposób jej względną częstotliwość, czyli prawdopodobieństwo pojawienia się w języku. Pamiętajmy o tym kiedykolwiek model języka wprawi nas w zachwyt swoją produkcją. Wyszukiwarka gotowców z naszego eksperymentu myślowego, równoważna inteligencją książce telefonicznej, posiada formalnie większą moc.

Pozostańmy jeszcze przez chwilę w sferze eksperymentów myślowych, gdzie mamy szanse poczynić kilka innych użytecznych obserwacji. Kreując hyperastronomicznych rozmiarów bazę tekstów równoległych na użytek tłumaczenia możemy rozrzutnie założyć, że zawiera ona wszystkie możliwe teksty języka źródłowego oraz odpowiadające im tłumaczenia. Dopóki zbiór tekstów jest skończony (do czego wystarczy jakiekolwiek ograniczenie długości pojedynczego tekstu), matematyk radośnie zgodzi się z nami, że eksperyment jest całkowicie legalny. Ważne jest jedynie by baza danych zawierała każdy tekst, jaki ktoś potencjalnie miałby ochotę tłumaczyć, być może włączając pozbawione sensu śmieci w rodzaju: „Stale defibrillator dilutes galaxies in pea soup” przetłumaczone jako „Stęchły defibrylator rozpuszcza galaktyki w grochówce”.

Matematycy często zbywają kwestię rozwiązalności problemu najprostszym (matematycznie najelegantszym) rozwiązaniem jakie przychodzi im do głowy nie troszcząc się o detale praktyczne, które być może uprościłyby (przyspieszyły) sam proces rozwiązania, gdyby jego realizacja była komuś do czegoś potrzebna. Skoro celem naszych rozważań było ustalenie czy problem da się (w ogóle) rozwiązać i udało nam się wykazać, że tak, to nie ma sensu zastanawiać się nad uproszczeniami rozwiązania, od czego są inżynierowie i rzemieślnicy. Zauważmy jednak, że modyfikując naszą bazę danych na użytek (pojedynczego) modelu języka, będziemy zainteresowani czymś innym niż w przypadku tłumaczenia, mianowicie reprezentacją sensu przez przechowywany w niej korpus. Chcemy, aby model języka oparty na tym korpusie produkował wysokiej jakości zdania, a mówiąc ściślej wartościował zdania zgodnie z ich sensem. Tak więc, produkując bazę danych dla takiego modelu musimy zatrudnić (oczywiście wyłącznie w myślach) ekspertów od języka i wiedzy, którzy wygenerują nam wszystkie sensowne zdania, a nie po prostu wszystkie zdania, jakie w ogóle da się formalnie utworzyć w języku. Nie zaburza to dopuszczalności naszego eksperymentu myślowego (zadanie jest formalnie wykonalne w skończonym czasie przy zatrudnieniu skończonej siły roboczej) wskazując jednocześnie na subiektywną rolę korpusu w modelu języka. Mówiąc inaczej, zestaw syntaktycznych przykładów zawartych w korpusie reprezentuje konkretną i potencjalnie głęboką wiedzę zakodowaną w statystykach występujących w nim napisów. To właśnie dlatego ChatGPT wypowiada się z sensem, pomimo że nic a nic z tego nie rozumie. Wynika stąd także, że gdybyśmy przypadkiem chcieli usprawnić nasz myślowy model tłumaczenia przez zredukowanie rozmiaru bazy danych, zaczęlibyśmy od języka docelowego, czyli od wstępnego wygenerowania wszystkich sensownych zdań w tym języku, a dopiero potem zajęli się kwestią przyporządkowywania im fraz języka źródłowego. Z której strony nie patrzeć na problem tłumaczenia, sprawny model języka docelowego jawi się tam jako podstawowy rekwizyt.

Zwróćmy jeszcze uwagę na pewną ogólną prawdę wyłaniającą się bokiem z naszej dygresji. Jeśli komputer wyprodukuje nam błyskotliwe rozwiązanie postawionego przed nim problemu, nie musimy od razu zachwycać się jego zdolnościami (a w szczególności inteligencją), gdyż może się okazać, że rzeczone rozwiązanie zostało po prostu przygotowane z góry i cichutko oczekiwało na nasze zapytanie. Gdy konstruktor Trurl przysłał Klapaucjuszowi swój najnowszy wynalazek, maszynę do spełniania życzeń,[31] wyposażył ją w artykuły, których, wedle jego przewidywań, kolega potrzebował do bieżących zajęć. Ponadto, antycypując dociekliwość Klapaucjusza, sam ukrył się we wnętrzu maszyny, tak więc, gdy tamten poprosił ją o kopię twórcy, Trurl mógł wyskoczyć na światło dzienne demonstrując potęgę swojego wynalazku oraz własną próżność. Jak pamiętamy, nie skończyło się to dla niego dobrze.

Powróćmy do rzeczywistości. Abstrahując od złożoności faktycznego algorytmu wyszukiwania optymalnych kontynuacji dla sekwencji zdań, powinniśmy wyraźnie zdać sobie sprawę, że jedynym istotnym wykładnikiem jakości całego procesu jest jakość funkcji wartościującej modelu rozumiana jako rzetelność produkowanego przezeń rankingu napisów. Rzetelność ta jest bez porównania trudniejsza do (praktycznego) osiągnięcia niż efektywny sposób jej wykorzystania do generowania poprawnych tekstów. To drugie można uzyskać na wiele sposobów i każdy średnio wprawny programista radośnie przystąpi do dzieła, jeśli tylko otrzyma rzetelny model języka jako gotowy moduł.

Na obecnym etapie naszej dyskusji, rola „gołego" modelu języka ma prawo być niejasna dla czytelnika. Jakiż jest sens w wartościowaniu zdań i zgadywaniu ich kontynuacji, skoro nie wiemy co chcemy powiedzieć? Pamiętajmy, że model języka ma stanowić wspólny element wszelkich zastosowań języka naturalnego i jego celem jest po prostu wartościowanie i produkowanie zdań poprawnych bez określonego celu. To trochę tak, jakbyśmy próbowali popisać się biegłością w języku przez improwizację, nie mając nic konkretnego do powiedzenia i nie wiedząc, co słuchacz chciałby usłyszeć. Chwytamy przypadkowe kawałki zdań i rozwijamy je dowolnie długo wtrącając, gdzie trzeba, znaki przestankowe i dbając jedynie o to, by generowane przez nas ciągi słów posiadały maksymalny (imponujący) sens z punktu widzenia statystyki języka. Nie będąc komputerami nie jesteśmy zwykle stawiani w podobnej sytuacji. Nasze procesy myślowe najprawdopodobniej nie przebiegają w ramach programów posklejanych z wyizolowanych algorytmicznych komponentów (modułów), które ktoś może nam kazać wykonywać bez związku z resztą naszych intelektualnych funkcji. Rola modelu języka stanie się jaśniejsza, gdy użyjemy go do rozwiązywania zadań lingwistycznych, których jednym z przykładów jest tłumaczenie. Zadanie takie ogranicza wybór w naszej zabawie generowania słów i zwrotów dostarczając modelowi treści (tematu) dla jego skądinąd niepohamowanej kreatywności. Celem modelu staje się wtedy wyszukanie najlepszej sekwencji, która spełnia ograniczenia (ramifikację) zadania. Zauważamy, że problemy takie jak znalezienie najlepszego szyku zdania (najlepszej permutacji słów w zdaniu), poprawnej interpunkcji (gdzie i czy wstawić przecinek) sprowadzają się do wartościowania sekwencji napisów; podpadają zatem pod ogólnie rozumianą funkcjonalność modelu.

Zdanie stanowi naturalną jednostkę tłumaczenia czy wypowiedzi, ale jak to doskonale rozumiemy, jego jakość (czy nawet interpretacja) może zależeć od kontekstu. Tak więc nie zakładamy z góry, że wartościowanie modelu języka ogranicza się do pojedynczych zdań. Może ono dotyczyć teoretycznie dowolnie długich sekwencji.

Jako nieco bardziej skomplikowany przykład rozważmy następujący fragment tekstu (początek zdania), który wyrwałem na chybił-trafił z jednego z dokumentów poniewierających się na pulpicie mojego laptopa: „W końcu to normalne, że jeśli ktoś chce coś …". Naszym celem jest ustalenie, w jaki sposób możemy to zdanie kontynuować. Wprowadzamy je jako dane do modelu języka, który na wyjściu dostarcza nam ranking opcji w postaci listy słów oraz ich wag (które można traktować jako

[31] Stanisław Lem, Cyberiada (Wielkie Lanie), Wydawnictwo Literackie, 2015.

prawdopodobieństwo), np. <zrobić: 0,015>, <kupić: 0,013>, <wypić: 0,007>, <sprzedać: 0,002>, itd. Teoretycznie, lista wyprodukowana przez model zawiera wszystkie możliwe słowa języka, z ich wszelkimi możliwymi odmianami i końcówkami. Oczekujemy oczywiście, że pewne słowa uznane będą za bardziej prawdopodobne niż inne; w szczególności, znakomita większość z nich (na przykład „samolot") uzyska znikomy ranking, tak że można je będzie całkowicie pominąć (podobnie jak „abakus" po słowie „Ala" – patrz Rysunek 6). Te które pozostaną uznamy za dopuszczalne warianty przedłużenia sekwencji o kolejne słowo a ich wagi odpowiadać będą częstotliwości, z jaką podobne przedłużenia pojawiają się w języku – według oszacowań modelu.

Znaki interpunkcyjne traktowane są jak słowa. Tak więc dla sekwencji „W końcu to normalne …" jedną z wysoko punktowanych opcji będzie zapewne przecinek (po którym, w następnej kolejności, prawdopodobnie pojawi się „że" jako jedna z głównych możliwości). Wystąpi tam pewnie także kropka, gdyż zakończenie zdania po „normalne" jest dopuszczalne, a przynajmniej taka możliwość nie jest wykluczona brakiem szerszego kontekstu. Piszemy „zapewne", „prawdopodobnie" i „być może", gdyż nie wiemy, jak funkcjonuje model (na obecnym etapie jest to dla nas czarna skrzynka) i nie jest jasne, jak wiele możemy od niego oczekiwać. Pamiętajmy, że wartościowanie sekwencji i opcji ich rozszerzenia dokonywane jest przez beznamiętny program (który ktoś musi mozolnie stworzyć), a nie przez nas – na podstawie naszego doświadczenia. Funkcjonalna definicja modelu nie wymusza automatycznie jego jakości formułując jedynie pewien postulat względem interfejsu oraz interpretacji wyników.

Zgadujemy, że wartościowanie napisów przez model dokonuje się na podstawie informacji wyekstrahowanej z korpusu. Jeśli korpus jest dostatecznie duży, możemy liczyć na to, że pewne zestawy słów pojawią się tam na tyle często, by model potrafił je uznać za poprawne i popularne. Na przykład, sekwencja „jeśli ktoś chce" występuje prawdopodobnie wiele razy w odpowiednio obszernym korpusie i jej kontekst może dostarczyć użytecznej informacji dla wartościowania słów i sekwencji, które mogą po niej nastąpić.

Uczyńmy w tym miejscu ważną obserwację. Pokusa by wbudowywać w model reguły gramatyczne, identyfikować słowa jako specyficzne części mowy lub zdania prowadzi na manowce, a przynajmniej nie wolno jej ulegać na samym początku zabawy z korpusem. Tak przebiegały pierwsze (i zdecydowanie nieudolne) próby mechanicznego generowania tekstów na użytek tłumaczenia.[32] Wystarczająco rozległy korpus zawiera w sobie dostatek informacji na temat gramatycznych faktów typu, że po „ona" następuje „poszła" a nie „poszedł", lub że rzeczownik „dżdżu" nie posiada mianownika (który w korpusie nigdzie się nie pojawia). Powracając (ostrożnie) do analogii z dzieckiem, dostrzeżemy, że jest ono w stanie opanować wiele zasad gramatyki, zanim (jeśli w ogóle) podda się formalnej edukacji w tym zakresie, wyłącznie na podstawie korpusu. Głównym powodem, dla którego studiowanie gramatyki jest ważne dla człowieka pragnącego osiągnąć biegłość w języku jest ograniczoność jego pamięci. Nie jesteśmy w stanie przechowywać w głowie pełnej informacji na temat wszystkich zasłyszanych przez nas wzorców poprawnego języka i przywoływać jej błyskawicznie na każde żądanie, dlatego wspomożenie naszej wiedzy zwięzłymi regułami posiada dla nas oczywisty utylitarny sens. Komputer nie cierpi na podobne przypadłości; wszelka użyteczna informacja jaką wyekstrahował z korpusu jest mu dostępna zawsze i natychmiast.

Trudno jednak oczekiwać, by realistyczny korpus (w odróżnieniu od bazy danych z naszego eksperymentu myślowego) zawierał wszystko, co da się sensownie wyrazić w języku. Im dłuższa sekwencja, tym mniejsza szansa, że wystąpi ona w korpusie literalnie, co powoduje, że wykrywanie kontekstu jest trudne. W powyższym przykładzie, próbując przewidzieć następnik słowa „coś", sami (jako ludzie) nie bardzo wiemy, co jest najlepsze, gdyż najzwyczajniej nie mamy pojęcia, o co chodzi. Tak więc nie

[32] William J. Hutchins, Machine Translation: Past, Present, Future. Chichester: Ellis Horwood, 1986.

obrazimy się na model, jeśli zaproponuje nam słowo „zrobić” jako najbardziej prawdopodobną kontynuację (która zresztą może być całkowicie słuszna i obiektywna). Wyobraźmy sobie teraz, że otrzymujemy na wejściu pełniejszą sekwencję obejmującą dwa zdania: „Finansowanie badań praktycznych przez podatnika ma sens mniejszy i wielce wątpliwy. W końcu to normalne, że jeśli ktoś chce coś …”. Nowe (poprzedzające) zdanie dorzucone do sekwencji sugeruje, że autor zamierza wyrazić obiekcję przeciwko wyrzucaniu pieniędzy w błoto w ramach jakiejś działalności naukowo-przemysłowej. Doprowadzi to prawdopodobnie do rozszerzenia naszych opcji o „wyprodukować” przesuwając „sprzedać” na wyższą pozycję i eliminując „wypić” jako kontynuację niedorzeczną w zaistniałym kontekście. Oczekiwalibyśmy czegoś podobnego od modelu, lecz nie zapominajmy, że posługuje się on jedynie korpusem a nie prawdziwym zrozumieniem treści.

Pierwsze modele języków oparte były na prostych zależnościach statystycznych wydobytych z niezbyt dużych korpusów składanych ręcznie. Korpus może nam pozwolić bezpośrednio identyfikować krótkie i popularne sekwencje przez zliczanie częstości ich literalnego wystąpienia i na tej podstawie nadawać wagi występujących w nim słowom. Kontekst tych wystąpień będzie jednak bardzo krótki lub bardzo dziurawy. Im dłuższy i pełniejszy kontekst, tym większa nadzieja, że ranking zostanie obliczony rzetelniej, jak w przypadku dwuzdaniowego tekstu z powyższego przykładu. Trudno jednak, poza sferą eksperymentów myślowych, liczyć na literalne (statystycznie częste) występowanie wszystkich interesujących nas długich fraz w korpusie.

Metody matematycznej statystyki pozwalają w pewnym zakresie wyciskać z modelu wskazówki, jak mechanicznie nadawać wagi sekwencjom, których reprezentacja w korpusie jest częściowa bądź przybliżona. Wyrażając się bardziej matematycznie, chodzi tu o obliczanie prawdopodobieństw warunkowych. Na przykład, porównując wartościowania sekwencji „mały piesek” oraz „mały pies” zapytamy, jak często w naszym korpusie pojawia się słowo „piesek” a jak często „pies”, jeśli poprzednim słowem jest „mały”. Obie sekwencje najprawdopodobniej wystąpią wiele razy w każdym przyzwoitym korpusie dostarczając sporo materiału do wartościowania, ale króciutki kontekst spowoduje, że wybór będzie posiadał znikome oparcie w treści analizowanej wypowiedzi. Jeśli naszym celem jest wygenerowanie następnika po słowie „mały” i wiemy (z postawionego przed nami zadania), że ma to być „piesek” albo „pies” (a nie na przykład „kot”), to nasz wybór między dwiema opcjami będzie siłą rzeczy przypadkowy. Ściśle mówiąc, wybierając opcję o wyższym wartościowaniu, wybierzemy po prostu to z dwóch dopuszczalnych słów, które w naszym korpusie częściej występuje po słowie „mały”. Jest to najlepszy możliwy wybór, jakiego potrafimy dokonać w danych warunkach.

Dłuższy kontekst, na przykład „Z jej skórzanej torby wyglądał mały …” pozwoli *nam* zgadnąć, że „piesek” jest prawdopodobnie lepszym wyborem, ale czy pozwoli także komputerowi? Szansa na literalne wystąpienie pełnej sekwencji w korpusie jest znikoma. Być może wystarczą ostatnie trzy słowa? Korpus może zawierać sekwencję: „skórzanej torby wyglądał mały piesek”, ale nie musi.

W takich warunkach przydają się heurystyki polegające na przymierzaniu korpusu w sposób przybliżony. Można na przykład policzyć jak często występuje w nim napis „mały piesek” poprzedzony słowem „torby”, być może odseparowanym od „mały” o kilka słów. Nie jest to precyzyjny algorytm, gdyż nie wiadomo na jakiej podstawie mamy pominąć słowo „skórzanej”. Nie jest jasne, czy powinno mieć ono wpływ na wybór, ale skąd niby komputer ma o tym wiedzieć? Z drugiej strony, może nie trzeba go pomijać i pozwolić statystyce zrobić swoje? Może „skórzane” torby są statystycznie mniejsze, więc wektor kontekstu będzie zwrócony w dobrą stronę?

Podobne dywagacje (a przytoczyliśmy jedynie znikomy fragment problematyki statystycznych modeli języka) sugerują, że korpus zawiera znacznie więcej użytecznej informacji niż da się dostrzec gołym okiem. Nie jest jednak jasne jak korzystać z tej informacji przy pomocy algorytmu, który musi się zdecydować na wymierne wagi, opcje, parametry – krótko mówiąc proste przesłanki dla interpretowania

treści korpusu w formie sekwencji symboli. Zalety sieci neuronowych, którymi się niebawem zajmiemy, okażą się kluczowe dla skutecznych prób wykorzystania ukrytej informacji drzemiącej w korpusie.

W przypadku modeli statystycznych, próby takie prowadzą do komplikowania modelu zawiłymi heurystykami (tolerowanie dziur między słowami, podpieranie korpusu regułami gramatycznymi) oraz parametrami numerycznymi (ważącymi wpływ heurystyk na wartościowanie), które można próbować korygować i mozolnie testować równolegle z powiększaniem (modyfikowaniem) korpusu. Weryfikowanie modelu przy takim podejściu musi być żmudne, gdyż ocena jego jakości dokonuje się na podstawie obserwacji przez człowieka, co wymaga analizy olbrzymiej liczby przykładów, oceny zdań produkowanych przez model (która jest często subiektywna) oraz mniej lub bardziej kabalistycznego odgadywania, w którym kierunku należy przesunąć parametry i poprawić heurystyki modelu by uzyskać lepszy wynik. Ranking wyprodukowany przez próby dopasowywania długich fraz wejściowych do dziurawej zawartości korpusu okazuje się często nieznacznie lepszy niż odgadywanie w ciemno i sporo prób ulepszenia modelu prowadzi na manowce. Taka była przeważająca metodologia postępu w budowaniu komputerowych modeli języka do końca pierwszej dekady obecnego stulecia,[33] zanim zatrudniono do tego sieci neuronowe.

Jeszcze trochę o tłumaczeniu

Jak zauważyliśmy wcześniej, założona funkcjonalność modelu języka umożliwia mu wspomaganie prostej gry towarzyskiej polegającej na tym, że ktoś proponuje jakieś słowo, lub sekwencję słów, a my potrafimy zasugerować kilka możliwych kontynuacji i formalnie (numerycznie) ocenić ich szanse okazania się najlepszymi rozszerzeniami z punktu widzenia statystyki języka reprezentowanego przez monojęzyczny korpus modelu. Ale to jeszcze nie wszystko co potrafimy. Nasz model proponuje zwykle więcej niż jedną kontynuację zadanej sekwencji, sugerując na ogół sporą liczbę opcji, których wartościowanie nie zawsze jest drastycznie zróżnicowane. Gdy patrzymy na podaną nam przez kogoś sekwencję słów, która wiąże się z jakimś zadaniem lingwistycznym, możemy ocenić jakość całej sekwencji jako funkcję jakości jej składników – słów lub większych fragmentów. Próby rozwiązania postawionego przed nami zadania mogą prowadzić do ograniczenia przestrzeni wyszukiwania dla modelu dostarczając w ten sposób tematu (treści) dla naszej gry.

Posiadając model języka docelowego A i chcąc go zastosować do tłumaczenia tekstów z języka źródłowego B na A, możemy rozbić proces tłumaczenia na dwa etapy. Pierwszą część procesu tłumaczenia zinterpretujemy jako identyfikację repertuaru słów oraz fraz języka A, z których budować będziemy tłumaczenie zadanego tekstu źródłowego, przy czym rzeczona identyfikacja nie musi być jednoznaczna i precyzyjna. Jej celem jest dostarczenie modelowi języka A dobrze postawionego zadania w postaci zestawu opcji do ewaluacji. Jego rozwiązaniem będzie poprawny (najwyżej wartościowany) tekst w języku A odpowiadający treścią źródłowemu tekstowi w języku B. Wyprodukowanie wynikowej formy takiego tekstu należy do kompetencji (monojęzycznego) modelu języka A.

Dla ilustracji, wyobraźmy sobie, że prosty moduł tłumaczący współpracujący z naszym modelem języka otrzymuje do przetłumaczenia angielską frazę „All that glitters isn't gold." Moduł dokonuje prostego słownikowego podstawienia słów pilnując jednak, by niczego nie zgubić. Jego produkt może wyglądać tak: „[wszystko, wszyscy, wszystkie] [co, że] [świeci, błyszczy, migoce] <się> nie <jest> [złoto, złote, złoty, złotem]". Tam, gdzie wybór konkretnej wersji tłumaczonego słowa wymaga kompetencji modelu języka A lub zrozumienia kontekstu, tłumacz umieszcza (w kwadratowych nawiasach) wszystkie możliwości (formy, przypadki) pobrane ze słownika. Tam, gdzie słowo może okazać się niepotrzebne (w języku

[33] Philipp Koehn, Statistical Machine Translation. Cambridge University Press, 2010.

angielskim świecenie nie jest zwrotne, lecz może się takim okazać w polskim tłumaczeniu, natomiast „jest" bywa pomijalne), ujęte jest ono w specjalne nawiasy reprezentujące opcjonalność.

Zadanie drugiej części procesu tłumaczenia jest teraz dobrze określone: używając modelu języka A należy wygenerować najlepsze zdanie pasujące do opcji dostarczonych przez moduł tłumaczący. Wolno nam próbować różnych kombinacji, włączając zmianę kolejności słów i poszukując takiej, która wyprodukuje najwyższe wartościowanie. Jeśli nasz model jest cokolwiek wart, możemy oczekiwać, że propozycja „Wszyscy jest nie że błyszczy złote" otrzyma zdecydowanie niższy ranking niż na przykład „Nie wszystko co się świeci jest złotem". Taki wynik można by od biedy uznać za poprawne tłumaczenie (szczególnie, że dokonał go komputer), ale nasz algorytm może jeszcze spróbować pominąć „jest" i skorzystać z interpunkcji, co powinno go doprowadzić do wersji wieńczącej dzieło, mianowicie: „Nie wszystko złoto, co się świeci", która ma spore szanse wystąpić literalnie w szanującym się korpusie. Rzecz jasna, w praktycznej implementacji postaramy się używać reguł, heurystyk oraz zaawansowanych technik algorytmicznych ograniczających liczbę bezsensownych prób. Nie jest to istotne dla zrozumienia zasady algorytmu. Jeśli wiadomo jaki jest cel i jakie są (skończone) możliwości manewru, usprawnianie procesu wyszukiwania to już zupełnie inny i znacznie łatwiejszy problem.

Poważne moduły tłumaczące nie funkcjonują na zasadzie słowników, nawet jeśli mogą liczyć na wysokiej jakości model A dla wypolerowania wyniku. Aby było co polerować, należy dobrze przygotować zadanie dla modelu A. Powyższy przykład nie oddaje pełnej złożoności problemu. Przede wszystkim, pojedyncze słowo nie zawsze jest właściwą jednostką tłumaczenia, podobnie jak nie zawsze pojedyncze słowo jest właściwym wskaźnikiem kontekstu na użytek modelu języka.

W przypadku tłumaczenia, podobnie jak w modelu monojęzycznym, kwestia formalnych reguł, a także słowników, schodzi na dalszy plan i pojawia się dopiero na etapie uzupełniających heurystyk. Informacja na temat tłumaczenia znajduje się w korpusie, choć sposób jej wykorzystania jest siłą rzeczy odmienny niż w przypadku modelu pojedynczego języka.

Głównym celem modułu tłumaczącego jest przyporządkowanie frazom języka źródłowego dopuszczalnych fraz języka docelowego, które mogą pojawić się w tłumaczeniu i których ostateczne potraktowanie dokona się przez model języka docelowego, na przykład według zgrubnego schematu opisanego powyżej. Otrzymawszy surowy korpus dwujęzyczny jako jedyne źródło wiedzy o relacji między językami B i A, moduł musi nauczyć się przyporządkowywać frazy (konstrukcje) języka B frazom języka A. W jaki sposób może on posiąść tę umiejętność? Wyobraźmy sobie, że korpus zawiera następującą parę tekstów (przykład tłumaczenia): „A little dog was lurking in her leather bag" oraz „Z jej skórzanej torby wyglądał mały piesek". Załóżmy, że moduł poszukuje opcji dla przetłumaczenia frazy „little dog", która występuje jako fragment aktualnie tłumaczonej sekwencji (pochodzącej z zupełnie innego kontekstu). Pojawienie się takiej frazy w pewnym elemencie korpusu oznacza jedynie tyle, że jest ona związana z jakąś sekwencją występującą w drugim elemencie pary. Która to może być sekwencja? Może „mały piesek"? Może „jej torebki"? Może „wyglądał mały"? Może „torebki piesek"? A może „jej wyglądał mały"? No bo przecież w tłumaczeniach tak bywa, że jedno słowo tłumaczy się na dwa, albo dwa na jedno, albo dwa na trzy, które na dodatek nie muszą występować w przetłumaczonej wersji jedno po drugim. Powtórzmy na czym polega problem. Nic nie rozumiemy i mamy do dyspozycji jedynie korpus. Naszym celem jest nauczyć się tłumaczenia. Jak pozyskać potrzebną nam wiedzę z przykładów, by dała się ekstrapolować na sekwencje, które w przykładach nie występują żywcem?

Widać, że jeden przykład to zdecydowanie za mało, ale posiadając ich odpowiednio wiele możemy zacząć polegać na statystyce. Jeśli bowiem znajdziemy w korpusie jakiś inny zestaw, którego pierwsza sekwencja zawiera frazę „little dog", powiedzmy „His little dog was very attached to him" przetłumaczoną jako „Jego mały piesek był do niego bardzo przywiązany", to nawet przy całkowitym braku zrozumienia znaczenia tych wszystkich symboli, potrafimy zauważyć, że „mały piesek" występuje w

tłumaczeniach z obu przykładów. Można z tego wydedukować całkiem mechanicznie, że ta fraza prawdopodobnie odpowiada „little dog" w tekście źródłowym. Piszemy „prawdopodobnie" a nie „na pewno", gdyż przytoczony przykład jest naiwny, a problem często bywa nieporównanie bardziej skomplikowany. Chcemy tylko pokazać w jaki sposób informacja na temat tłumaczenia poszczególnych słów i fraz może być wydobyta z korpusu i skwantyfikowana przy pomocy miar statystycznych bez cienia próby zrozumienia treści tłumaczonych fraz. Mechanizmy jej kwantyfikowania pozwolą procesowi tłumaczącemu wartościować różne kandydatury na poprawne tłumaczenia, podobnie jak model monojęzyczny potrafi wartościować zdania z języka docelowego. Procedurę wartościowania tych kandydatur nazwiemy modelem tłumaczenia.

Warto się nad tym przez chwilę zastanowić. Gdybym posiadał olbrzymią i absolutnie rzetelną pamięć, gdybym potrafił błyskawicznie wyszukiwać w niej symboliczne frazy i gdybym otrzymał odpowiednio rozległy korpus, to nauczyłbym się tłumaczyć z hebrajskiego na chiński bez śladu zrozumienia treści najprostszych zdań w którymkolwiek z tych języków. Moja aktywność w procesie tłumaczenia stanowiłaby równoważnik przeszukiwania kartoteki z szufladkami indeksowanymi niezrozumiałymi obrazkami, zliczaniu obrazków na przeróżnych kartkach wydobytych z rzeczonych szufladek, kopiowaniu obrazków na inne kartki na podstawie wyników moich obliczeń i tak dalej. Powrócimy jeszcze do tej scenki.

Mamy zatem dwa modele: języka docelowego A oraz tłumaczenia z języka B na język A. Rysuje się pewien schemat zaprzęgnięcia obu modeli do wspólnego zadania. Model tłumaczenia generuje i wartościuje zestawy możliwych tłumaczeń na poziomie fraz dopasowanych w oparciu o statystyki korpusu dwujęzycznego, z których model języka docelowego (oparty na korpusie monojęzycznym) buduje i wartościuje poprawne sekwencje wynikowe. Celem jest wyprodukowanie najlepszego tłumaczenia w oparciu o wspólne wartościowanie traktowane jako kryterium jakości wyniku. Ta pojedyncza wartość jest funkcją obu wartościowań, gdzie mogą wchodzić w grę subtelne heurystyki, jak na przykład odległość tłumaczonych fraz w zdaniu docelowym. Przy porównywalnej ocenie kilku wariantów tłumaczenia wyprodukowanych przez model języka A wybieramy ten, w którym odległości odpowiednich fraz są bardziej zbliżone do ich odległości w oryginale.

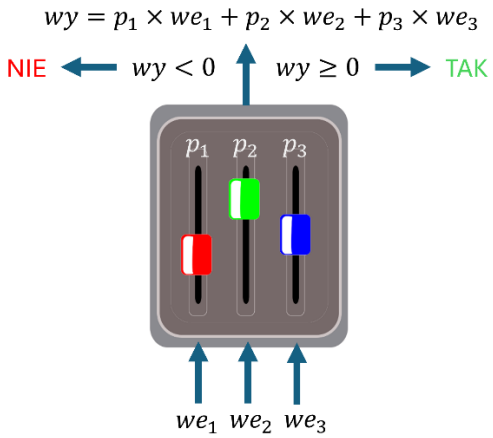
Statystyczne podejście do tworzenia modeli języków i tłumaczenia stanowiło podstawę praktycznych wysiłków w tym zakresie mniej więcej do roku 2015. Najpopularniejsza platforma automatycznego tłumaczenia, Google Translate,[34] bazowała do roku 2016 na metodach statystycznych. We wrześniu 2016, Google zapowiedział (zrealizowane w listopadzie tegoż roku) przejście na system oparty na sieciach neuronowych, którą to datę można uznać za moment przełomowy rozpoczynający nową erę w modelowaniu języków naturalnych.

Sieci neuronowe

Sieć neuronowa, podobnie jak model statystyczny, może posłużyć do reprezentacji pewnej niekoniecznie precyzyjnej wiedzy drzemiącej w zbiorze przykładów (czyli korpusie), gdzie celem jest umiejętność klasyfikacji, czyli wartościowania opcji i dokonywania „inteligentnego" wyboru. W porównaniu z modelem statystycznym, sieć neuronowa posiada co najmniej dwie zalety. Po pierwsze, z samej swojej natury, wyprodukuje ona zawsze jakiś wynik, co znaczy, że twórca modelu nie musi się martwić, co robić, gdy korpus nie zawiera danych pozwalających bezpośrednio wartościować konkretny przypadek. Nie ma oczywiście żadnej apriorycznej gwarancji jakości wyniku, szczególnie jeśli korpus posiada poważne dziury, lecz istnieją za to naturalne metody poprawiania tej jakości nie wymagające tworzenia pokrętnych heurystyk i zgadywania ich parametrów przez przyglądanie się fusom po kawie. Na tym zasadza się druga wielka zaleta sieci neuronowych: potrafią one się uczyć lub – jak kto woli – trenować. Istnieje pewien systematyczny proces (algorytm), według którego sieć neuronowa może poprawiać

[34] https://translate.google.com/ .

jakość swojej funkcji w oparciu o wyniki testów. Proces ten może przebiegać prawie całkowicie automatycznie nie wymagając ludzkiej analizy ani nawet zbyt wnikliwej interwencji. Ceną, którą za to płacimy jest brak dokładnego zrozumienia co się dzieje we wnętrznościach sieci neuronowej, szczególnie jeśli jest to sieć duża i wyuczona na wielkim korpusie. Nadaje to jej pozór cech ludzkich w postaci dozy nieprzewidywalności i omylności. O ile w przypadku modelu statystycznego, odpowiedzialność za jego niedoskonałości można zrzucić na twórcę modelu, który musiał dostarczyć algorytmów i formuł dających się uargumentować matematycznie, o tyle w przypadku dostatecznie skomplikowanej sieci neuronowej nie jesteśmy w stanie szczegółowo wytłumaczyć jej pomyłek. Nie chcę tu powiedzieć, że są to obiekty tajemnicze i nieznane. Wręcz przeciwnie – matematyka sieci neuronowych jest dobrze rozwinięta i w znacznej części dostępna nawet dla początkujących adeptów. Mechanizm ich funkcjonowania maskuje jednak skutecznie detale wpływające na poszczególne przypadki wartościowania i klasyfikacji. Tak musi być. Albo sami żmudnie wytyczamy algorytmiczne ścieżki interpretowania korpusu (ograniczając model regułami, które jesteśmy w stanie ogarnąć), albo zdajemy się na proces samoczynnego uczenia się sieci, która automatycznie czerpie ze złożoności informacji zawartej w korpusie budując ukradkiem swoje własne, nieznane nam reguły.

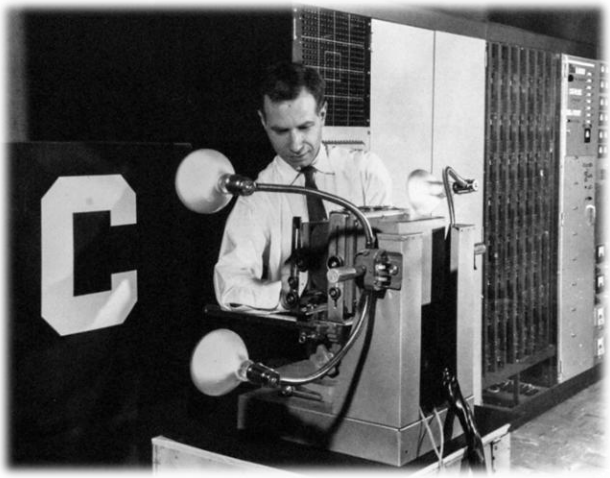


Rysunek 7. Perceptron z trzema wejściami.

Jak wskazuje nazwa, sieć neuronowa jest układem złożonym z połączonych ze sobą elementów zwanych neuronami. Idea datuje się od roku 1943, w którym pojawił się artykuł proponujący matematyczny model aktywności biologicznych neuronów.[35] Model, nazwany później perceptronem, zakładał, że pojedynczy neuron funkcjonuje jako wyzwalacz sterowany prostą funkcją impulsów nadchodzących z synaps.

W swojej realizacji, perceptron stanowił rodzaj prościutkiego miksera sygnałów przysposobionego do rozwiązywania zadania klasyfikacji (Rysunek 7). Poziomy sygnałów elektrycznych pojawiających się na określonej liczbie wejść miksera regulowane były potencjometrami (w sposób podobny do współczesnego miksera akustycznego) i kierowane na wspólne wyjście, gdzie ich wypadkowa podlegała ocenie TAK lub NIE, w zależności od tego, czy sygnał (mówimy tu o sumarycznym napięciu) był dodatni czy ujemny. Urządzenie było zatem nieporównanie prostsze od miksera akustycznego, gdyż interesowało się wyłącznie prądem stałym. Jego związek z funkcją biologicznego neuronu sugerował powiązania z procesami myślowymi zachodzącymi w ludzkim mózgu. Warto zauważyć, że dziś, kiedy zastosowania perceptronu zaczynają wkraczać na teren skutecznego imitowania intelektualnej aktywności człowieka, tego typu konotacje formułowane są ostrożniej niż w owych pionierskich czasach.

[35] Warren McCulloch and Walter Pitts, A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 1943.



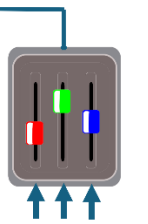
Rysunek 8. Frank Rosenblatt i Mark 1 (Wikimedia Commons).

Idea perceptronu w zastosowaniu do klasyfikacji sygnałów (z gruntu analogowych) sugerowała budowanie maszyn przetwarzających rzeczone sygnały w oparciu o zasady odmienne od zasad (cyfrowych) komputerów. Pamiętajmy, że w owych czasach praktycznie wszystkie interesujące dane (sygnały) posiadały reprezentację wyłącznie analogową (radio, telefonia, telewizja, prasa, książki, fotografie). Pomimo wczesnych eksperymentów z programowanymi implementacjami perceptronów na zwykłych cyfrowych komputerach prowadzonych przez IBM, pierwsza próba ich poważnego zastosowania odbyła się z wykorzystaniem specjalistycznego sprzętu (komputera analogowego) o nazwie Mark 1 (Rysunek 8). Projekt przeprowadzony pod koniec lat 50-tych w Cornell Aeronautical Laboratory i finansowany przez marynarkę wojenną USA posiadał charakter wojskowy, a jego celem była automatyczna klasyfikacja obrazów. Konferencja prasowa zorganizowana na jego temat w roku 1958 przyczyniła się do radosnego nagłośnienia nowego narzędzia w mediach,[36] które z perspektywy lat zakrawa na (w znacznej mierze skuteczną) kontrwywiadowczą kampanię dezinformacyjną. Czytelnicy dowiedzieli się mianowicie, że marynarka wojenna posiadła właśnie nową wersję komputera, która niebawem nauczy się chodzić, widzieć, słyszeć, czytać, pisać, rozmawiać ludzkim głosem, takoż reprodukować się, a na dodatek pozyska świadomość istnienia. Euforię prasowych doniesień podsycał fakt, że perceptron był w stanie sam się uczyć. Ustawienia potencjometrów miksujących sygnały wejściowe dawało się bowiem korygować automatycznie w oparciu o różnicę między sygnałem wyjściowym a jego oczekiwaną wartością. Jeśli urządzeniu dostarczono zestaw przykładów zawierający konfiguracje sygnałów wejściowych z ich wzorcową (poprawną) klasyfikacją, mogło ono samo ustawić potencjometry w ten sposób, by zredukować zakres błędów. Procedura posiadała charakter iteracyjny; nauczanie perceptronu przebiegało metodą kolejnych przybliżeń. Startując z pewnego (dowolnego) początkowego ustawienia potencjometrów, system przeglądał przykłady, dokonywał ich klasyfikacji i oceniał (porównując sygnały wyjściowe i ich oczekiwanymi wartościami), jak dalece otrzymane wyniki różnią się od poprawnych. Rozbieżności tłumaczyły się w wartości liczbowe wskazujące o ile ustawienia poszczególnych potencjometrów należy zmienić, by próbować zmniejszyć błąd w następnej rundzie.

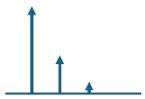
Spróbujmy nabyć odrobinę intuicji na temat tej operacji. Warto się nad tym zastanowić, gdyż 95% finezji wszystkich współczesnych sieci neuronowych zasadza się na prostej obserwacji, którą uczynimy za moment. Jeśli dla danego zestawu sygnałów perceptron wyprodukuje odpowiedź TAK, podczas gdy poprawna odpowiedź brzmi NIE (jak dla przykładu numer 2 z Rysunku 9), oznacza to, że należy odrobinę zredukować ustawienia potencjometrów, czyli przesunąć je w dół. Poziom sygnału na wyjściu jest

[36] Mikel Olazaran, A Sociological Study of the Official History of the Perceptrons Controversy. Social Studies of Science, 26(3), pp. 611-659, 1996.

bowiem zbyt wysoki dla danej konfiguracji sygnałów wejściowych. Na tej samej zasadzie, jeśli poprawnym wynikiem jest TAK a uzyskaliśmy NIE (przykład numer 7), wypadnie potencjometry popchnąć do góry. Na ile to możliwe, należałoby postępować według jakiegoś rozsądnego schematu starając się unikać sytuacji, w której naprawiając wynik dla jednego przykładu psujemy go dla innych.



przykład	we_1	we_2	we_3	oczekiwany	uzyskany
1	9	7	4	TAK	TAK
2	4	9	8	NIE	TAK
3	8	6	4	NIE	NIE
4	6	2	1	TAK	TAK
5	5	6	1	NIE	NIE
6	7	9	3	NIE	NIE
7	9	4	1	TAK	NIE



Rysunek 9. Zasada nauczania perceptronu.

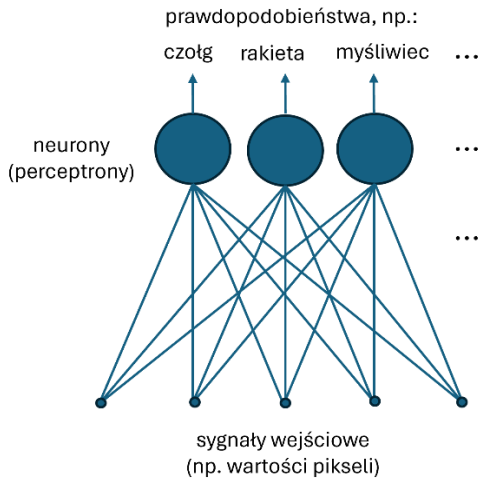
Popatrzmy uważnie na wartości sygnałów wejściowych w zestawie numer 7 na Rysunku 9 i zastanówmy się, który z trzech potencjometrów posiada największy wpływ na poziom sygnału wyjściowego. Jest to oczywiście ten potencjometr, dla którego poziom wejściowego sygnału jest największy, czyli potencjometr numer 1. Gdyby nam przypadkiem przyszło do głowy użyć potencjometru numer 3 dla podbicia poziomu sygnału wyjściowego (dostatecznego dla zmiany niepoprawnej klasyfikacji), musielibyśmy przesunąć go znacznie dalej niż potencjometr numer 1 ryzykując w ten sposób rozkalibrowanie perceptronu dla pozostałych zestawów. Naszym celem jest więc maksymalizacja wzrostu poziomu sygnału wyjściowego przy minimalnym totalnym zaburzeniu ustawienia potencjometrów. Matematyczna konkluzja jest taka: potencjometry należy przesuwać proporcjonalnie do siły odpowiadających im sygnałów wejściowych. Formalnie oznacza to tyle, że zakres zmiany ustawienia potencjometru numer i ma być równy iloczynowi $F \times we_i$, gdzie F jest stałym współczynnikiem, a we_i jest poziomem sygnału wejściowego numer i. Kierunek zmiany odpowiada różnicy między odpowiedzią oczekiwaną a wyprodukowaną przez perceptron.[37]

Współczynnik F jest jednym z parametrów procedury uczenia ustalonym przez eksperymentatora trenującego perceptron i opisującym gwałtowność zmian w pojedynczym kroku. Większa wartość F może przyśpieszyć trening, ale może także powodować przestrzelenia i oscylacje uniemożliwiające precyzyjne dostrojenie perceptronu do danych. Innym parametrem procedury uczenia jest kryterium zakończenia procesu, czyli uznania, że dalsze próby nie mają sensu, gdyż nie poprawią już więcej wyników. Jasne, że błędy nie zawsze dają się całkowicie wyeliminować, ponieważ perceptron nie jest w stanie poprawnie klasyfikować wszystkiego. Niektóre błędy można przyjąć za akceptowalne; czasem należy skapitulować i stwierdzić, że urządzenie nie potrafi należycie rozwiązać postawionego przed nim zadania.

Dzisiejsze sieci neuronowe (perceptronowe) implementowane są praktycznie wyłącznie jako algorytmy komputerowe, co w znacznej mierze wynika z powszechnej cyfryzacji wszelkich rodzajów informacji. Tworzone są oczywiście specjalizowane komputery (procesory) pozwalające implementować sieci

[37] Zauważmy, jeśli nie jest to oczywiste z naszej dyskusji, że wagi p_i reprezentowane przez nasze umowne potencjometry mogą być ujemne.

neuronowe znacznie efektywniej niż tradycyjne komputery, nie są to jednak zestawy potencjometrów pokręcanych przez motory i nie operują one na sygnałach analogowych. Zmieńmy zatem terminologię na bardziej współczesną. Perceptron staje się neuronem, a potencjometr przeobraża się w liczbowy współczynnik, który nazwiemy wagą wejściowego sygnału. No a sygnał to po prostu liczba – jak to w komputerze. Cała reszta pozostaje bez zmian, z wyjątkiem drobnego szczegółu: funkcja wyjściowa nie musi być interpretowana jako prosta binarna klasyfikacja (TAK lub NIE). Na wyjściu sygnał wolno przekształcić (posługując się funkcją z określonego repertuaru) i na przykład potraktować go jako ciągłe prawdopodobieństwo.

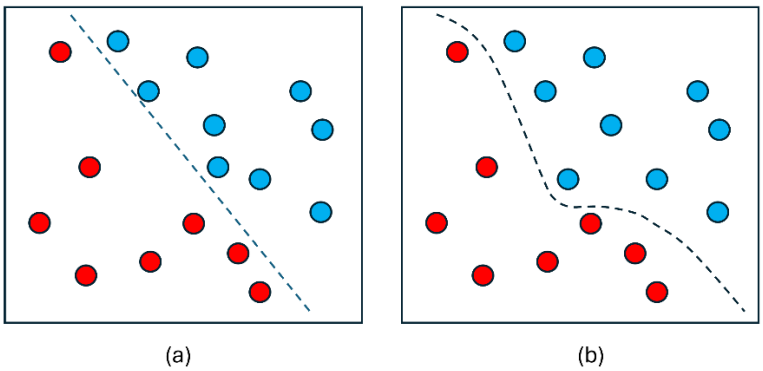


Rysunek 10. Schemat sieci neuronowej (perceptronowej) do klasyfikacji obrazów.

Łatwo dało się zauważyć, że zestaw neuronów (perceptronów) można połączyć w grupę – w ten sposób, że używając tych samych wejść grupa wyprodukuje bardziej skomplikowane wyjście składające się z kilku sygnałów. Tak właśnie funkcjonował Mark 1, którego celem, jak pamiętamy, było rozpoznawanie obrazów. Połączenie wejść w grupie jest równoległe (patrz Rysunek 10). Te same sygnały wejściowe powielone są na wejściach wielu neuronów, z których każdy używa swoich prywatnych ustawień potencjometrów-wag produkując osobny sygnał wyjściowy dla danej konfiguracji sygnałów na wejściu (dokonując niejako osobnej i indywidualnej klasyfikacji). Bardziej naukowo powiemy, że układ tego typu zamienia wektor sygnałów wejściowych na wektor sygnałów wyjściowych, przy czym rozmiary tych wektorów mogą być różne. Liczba elementów wektora wejściowego równa jest przyjętej liczbie sygnałów wejściowych (takiej samej dla każdego neuronu), natomiast liczba sygnałów wyjściowych równa jest liczbie neuronów w zestawie. Wyobraźmy sobie (cokolwiek przeceniając możliwości takiego układu), że faktycznie chcemy go zastosować do rozpoznawania obrazów. Sygnały wejściowe reprezentują zatem wartości pikseli pobranych z obrazka, natomiast sygnały wyjściowe odpowiadają typom klasyfikacji. Pozostając w duchu historycznego projektu, możemy sobie wyobrazić, że pierwszy sygnał wyjściowy to czołg, drugi to rakieta, trzeci to myśliwiec, itd. Na wejściu podstawiamy obrazek reprezentowany wartościami pikseli, a na wyjściu pojawia się liczba odpowiadająca prawdopodobieństwu, że obrazek zawiera określony obiekt. Pojedynczy neuron zajmuje się więc klasyfikacją określonego typu obiektów. Decydując się na wynik globalnej klasyfikacji wybieramy wyjście o największej wartości.

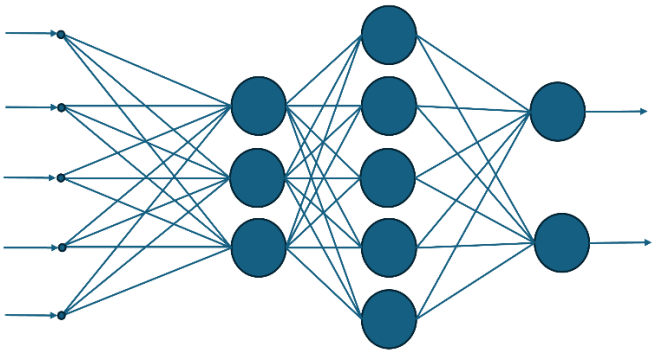
To co opisaliśmy powyżej nazywa się jednowarstwową siecią neuronową. Jej możliwości są bardzo ograniczone, gdyż zakres funkcji dostępnych pojedynczemu neuronowi jest skromny. Wprawdzie potrafi on samoczynnie redukować błąd klasyfikacji sygnałów wejściowych w oparciu o przykłady, ale istnieją proste problemy, których poprawnej klasyfikacji neuron nigdy się nauczy. Wyobraźmy sobie, że otrzymujemy kartkę z zaznaczonymi na niej punktami, przy czym punkt może być albo czerwony, albo niebieski (patrz Rysunek 11). Neuron otrzymuje współrzędne tych punktów (jako pary sygnałów wejściowych) oraz informację o kolorze punktu i chce się nauczyć, który punkt jest czerwony a który

niebieski, tak by odtąd, otrzymawszy jedynie współrzędne, mógł podać kolor punktu bez spoglądania na kartkę. Będzie to możliwe jedynie wtedy, gdy punkty na kartce da się oddzielić linią prostą, tak by po jednej stronie linii znajdowały się wyłącznie punkty o tym samym kolorze. W przeciwnym przypadku, klasyfikacja punktów przez neuron musi zawierać błędy niezależnie od ustawień wag. Mówiąc matematycznie, funkcja obliczana przez pojedynczy neuron musi być funkcją liniową.



Rysunek 11. Przykłady prostych (decyzyjnych) problemów klasyfikacyjnych: (a) problem rozwiązywalny przy pomocy pojedynczego neuronu; (b) problem nierozwiązywalny przez pojedynczy neuron.

Po stosunkowo krótkiej euforii, nastąpiło zatem rozczarowanie. Głośna i kontrowersyjna (szczególnie z późniejszej perspektywy) książka dwóch ówczesnych luminarzy sztucznej inteligencji[38] zniechęciła sporą część badaczy do idei perceptronu-neuronu demonstrując czarno na białym, że mizerny zakres funkcji jakie jest on w stanie realizować nie przystaje do ambitnych wizji nagłaśnianych przez prasę. Ich autorytatywna konkluzja była taka, że rozwój sztucznej inteligencji powinien przebiegać inną drogą. Jedną z pobocznych konsekwencji ówczesnych wysiłków popularyzatorskich było uznanie Franka Rosenblatta, szefa projektu w Cornell Aeronautical Laboratory, za wynalazcę perceptronu. Nie jest to jedyny wynalazek w historii postępu, którego ojcostwo (czy macierzyństwo) przypisane zostało mylnie przez media.



Rysunek 12. Przykład trójwarstwowej sieci neuronowej.

Fakt, że warstwa neuronów, jak ta z powyższego przykładu, posiada wiele wejść i wiele wyjść (i że wyjścia to też sygnały), sugeruje tworzenie sieci neuronowych z wielu warstw, w ten sposób, że wyjścia niższej warstwy stanowią wejścia warstwy wyższej (Rysunek 12). Pomysł ten przyszedł oczywiście do głowy wielu badaczom jako nadzieja na pokonanie dotkliwych ograniczeń pojedynczej warstwy neuronów. Okazało się jednak, że algorytm uczenia (trenowania) sieci, prosty i naturalny w przypadku jednej warstwy, nie jest bynajmniej oczywisty w przypadku większej liczby warstw. Mówiąc bardziej precyzyjnie, algorytm jako taki był wprawdzie stosunkowo prosty do uogólnienia, lecz intuicyjnie wydawał się

[38] Marvin Minsky and Seymour A. Papert, Perceptrons, an Introduction to Computational Geometry. MIT Press, Cambridge MA, 1969.

mało skuteczny. Niewątpliwy wpływ na pesymistyczne potraktowanie problemu przez społeczność badaczy odegrała negatywna i prewencyjna opinia prominentów (patrz wyżej). Rzeczona społeczność zaniechała zatem wysiłków w temacie sieci wielowarstwowych dochodząc do (jak się później okazało przedwczesnego) wniosku, że są one z przyrodzenia tępe i nie pozwolą się uczyć. Pewną rolę odegrał też fakt, że budowanie i emulowanie skomplikowanych sieci neuronowych na ówczesnych komputerach nie było ani łatwe, ani szybkie, ani przyjemne.

Stagnacja trwała do roku 1986, kiedy to pojawił się przełomowy artykuł wykazujący skuteczność i atrakcyjność tak zwanego algorytmu wstecznej propagacji,[39] który utorował ścieżkę dla efektywnego trenowania sieci wielowarstwowych. Od tamtego momentu zaznacza się wzrost zainteresowania wielowarstwowymi sieciami neuronowymi nasilający się w miarę rozwoju Internetu. Łatwa dostępność olbrzymich (i coraz większych) ilości danych stymulowała próby ich klasyfikacji w celach jak najbardziej praktycznych: rozpoznawanie mowy, obrazów, tekstów pisanych, oczywiście modelowanie języka i tłumaczenie, oraz wiele innych. Rozległe zasoby materiałów treningowych dla sieci neuronowych dostępne w Internecie ułatwiały i napędzały kreatywność przez utworzenie atmosfery kompetycji. Różne grupy badawcze konkurowały między sobą produkując coraz lepsze klasyfikatory. Jakość ich produktów była jak najbardziej wymierna, gdyż stopa błędów klasyfikatora obserwowana na probierczym zestawie przykładów jest prostym numerycznym wskaźnikiem jego jakości. Współzawodnictwo posiadało zatem wyraźne, jednoznaczne i obiektywne kryteria.

Neuronowe modele języka

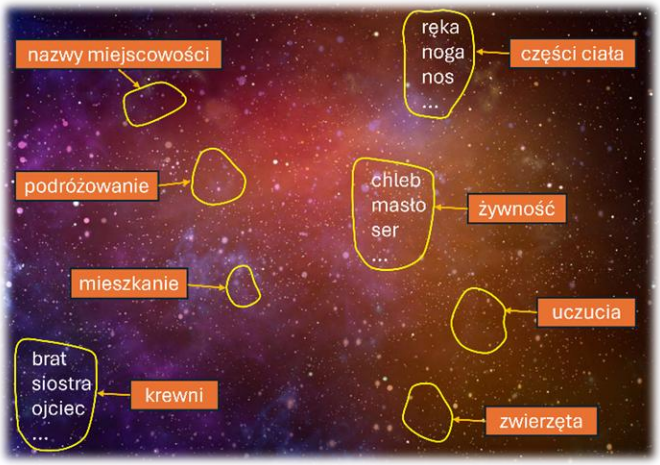
Wykorzystanie sieci neuronowych do budowania modeli języka wymagało wstępnego rozwiązania kilku problemów. Od samego początku było jasne, że muszą to być sieci olbrzymie. Tak więc pierwszym wymogiem sukcesu była możliwość efektywnego tworzenia bardzo dużych sieci (o wielkiej liczbie neuronów i wag), których trening dałby się przeprowadzić w rozsądnym czasie. Były do tego potrzebne duże i szybkie komputery wyposażone w wielką pamięć oraz procesory wspierające pewne typy obliczeń, mianowicie obliczenia wektorowe. Zarówno bowiem nauczanie sieci, jak i korzystanie z jej wyuczonej wiedzy, wymaga wielu identycznych (i niezależnych) operacji arytmetycznych wykonywanych jednocześnie na wszystkich wejściach i wagach neuronów danego poziomu. Jest to więc doskonały przykład problemu, gdzie umiejętność masywnego zrównoleglenia obliczeń przynosi bezpośredni zysk czasowy wprost proporcjonalny do ilości dostępnych zasobów. Typem procesora odpowiednim do zastosowania w implementacji sieci neuronowych są karty graficzne, które zostały zaprojektowane w celu szybkiego przetwarzania zawartości ekranów składających się z wielu milionów pikseli, gdzie częstym zadaniem jest zastosowanie tej samej operacji do dużego fragmentu obrazu. Poważne (komercyjne) sieci neuronowe budowane są w oparciu o dedykowane procesory (tzw. procesory tensorowe) funkcjonujące podobnie do kart graficznych, lecz skonstruowane specjalnie do zastosowań w sieciach neuronowych.

Aby zrozumieć rozmiar problemu, a przy okazji docenić tempo rozwoju technologii raz uznanych za komercyjnie atrakcyjne, po czym pośpiesznie doinwestowanych, porównajmy liczbę parametrów dwóch ostatnich wersji GPT,[40] czyli modeli języka napędzających ChatGPT, mianowicie wersji 3 oraz 4. Intuicyjnie, liczba parametrów modelu (w tym przypadku sieci neuronowej) odpowiada liczbie wag (koncepcyjnych potencjometrów), których ustawienia dokonują się w trakcie trenowania sieci. Dla wersji 3.5, liczba ta wynosi około 175 miliardów (175,000,000,000), podczas gdy wersja 4 posiada ponad

[39] Hinton Rumelhart, Learning Representations by Back-propagating Errors. Nature, 323 (6088), pp. 533-536, 1986.

[40] GPT jest skrótem od „Generative Pre-trained Transformer" i stanowi model języka (w sensie naszej dyskusji). ChatGPT jest w tym kontekście jedną z aplikacji tego modelu.

bilion (1,000,000,000,000) parametrów.[41] Jedną z zalet sieci neuronowych jest względna łatwość ich powiększania. Ideowo prosta struktura połączeń sprawia, że jedyny istotny problem stanowi koszt; także zakres dostępnej technologii wynika prawie bezpośrednio z zakresu inwestycji. Dla porównania, połączenie tysiąca laptopów w konglomerat użyteczny z punktu widzenia praktycznej mocy obliczeniowej na użytek rozwiązywania problemów numerycznych (jak na przykład modelowanie klimatu) nie jest pomysłem godnym uwagi. Połączenie tysiąca procesorów tensorowych, wykorzystywanych w sieciach neuronowych, jest natomiast stosunkowo proste, naturalne i jak najbardziej korzystne.



Rysunek 13. Idea zanurzania słów i fraz polega na umieszczaniu ich w pewnej abstrakcyjnej przestrzeni o bardzo dużej liczbie wymiarów w ten sposób, że semantycznie powiązane ze sobą słowa okupują wspólne rejony.

Jednym z fundamentalnych problemów na drodze do urzeczywistnienia neuronowych modeli języka był sposób przeobrażenia napisów w sygnały. Problem występuje po obu stronach sieci neuronowej, gdyż poza zamianą wejściowej sekwencji słów na wartości liczbowe przedłożone pierwszej warstwie neuronów, model musi jeszcze zamienić na ciągi słów także sygnały wyjściowe. Ten drugi etap jest jednak (przynajmniej koncepcyjnie) prostszy, gdyż – jak pamiętamy – formalnym celem modelu jest produkowanie prawdopodobieństw, czyli liczb, co neurony potrafią czynić bezpośrednio. Istnieje wiele mniej lub bardziej naturalnych z punktu widzenia komputera sposobów reprezentacji tekstów w postaci sekwencji bitów, a więc liczb, jednak intencją tych wszystkich historycznych schematów było proste kodowanie na użytek przechowania tekstu w pamięci komputera lub przesyłania go na odległość, a nie traktowanie ich jako sygnałów podlegających interpretacji w kontekście inteligentnego korelowania z innymi napisami. Tak więc na przykład numeryczna reprezentacja słowa „duży” nie ma nic wspólnego z reprezentacją słowa „wielki”. Przy transformacji tekstu dla sieci neuronowej warto było zatroszczyć się o to, by słowa czy zwroty bliskoznaczne lub pochodzące z tego samego zakresu tematycznego (jak „pies” i „kot”), posiadały reprezentacje, które w jakiś sposób znajdują się blisko siebie w sensie matematycznym, czyli wspierającym efektywne trenowanie modelu. Silną stroną sieci neuronowych jest bowiem wychwytywanie schematów i szablonów, co trudno czynić, gdy sygnały odpowiadające podobnym (lub przeciwnym) elementom wejściowym przydzielane są przypadkowo i nie zawierają liczbowych zależności korelujących się z ich semantycznymi powiązaniami. Badania w tym zakresie doprowadziły do idei zanurzania,[42] czyli reprezentowania słów jako punktów w pewnej wielowymiarowej

[41] Warto zdawać sobie sprawę z różnic w nazewnictwie dużych liczb pomiędzy systemem amerykańskim a brytyjskim (używanym także w Polsce). Łatwo tu o pomyłkę, szczególnie przy bezkrytycznym wspieraniu się informacją pobieraną z Internetu. Dla przykładu, polski miliard to amerykański „billion”.

[42] Ang. „embedding”, np. Yang Li and Tao Yang, Word Embedding for Understanding Natural Language, a Survey. Guide to big data applications, pp. 83-184, 2018.

przestrzeni, gdzie wymierna odległość między punktami wiąże się z semantyczną odległością między odpowiadającymi im słowami.[43]

Klasyczna sieć neuronowa, odpowiadająca opisanej wcześniej naturalnej generalizacji perceptronu, posiada określoną i ustaloną liczbę wejść i wyjść (Rysunek 12), która nie zmienia się w zależności od specyficznego problemu, jaki zostaje sieci przedstawiony. Jej funkcja polega na przyjęciu kompletnego zestawu sygnałów z wejścia i wyprodukowaniu wartości wyjściowych, co stanowi rozwiązanie jednej pełnej instancji problemu. W przypadku zastosowania takiej sieci do modelu języka oznaczałoby to tyle, że każda sekwencja słów pojawiająca się na wejściu modelu musiałaby składać się z tej samej, z góry ustalonej liczby elementów (czyli słów zamienionych na sygnały) powodując wygenerowanie określonego (pojedynczego) wektora prawdopodobieństw na wyjściu, gdzie pojedyncza wartość wyjściowa określałaby ranking określonego słowa ze słownika. Tymczasem, naturalnym i oczekiwanym rozwiązaniem jest takie, przy którym tekst wejściowy może posiadać dowolną długość, a jego prezentowanie modelowi odbywa się stopniowo, w miarę pojawiania się nowych słów czy fraz. W ten sposób model mógłby posłużyć do prowadzenia konwersacji akceptując kolejne frazy od rozmówcy i produkując wynikające z nich odpowiedzi. Taka jest właśnie zasada modeli generatywnych, których przedstawicielem jest ChatGPT. W miarę narastania wejścia, model coraz lepiej „wczuwa się" w temat. Zarówno cała sekwencja pochodząca od rozmówcy jak i wygenerowana dotąd odpowiedź służą generatorowi jako parametryzacja.

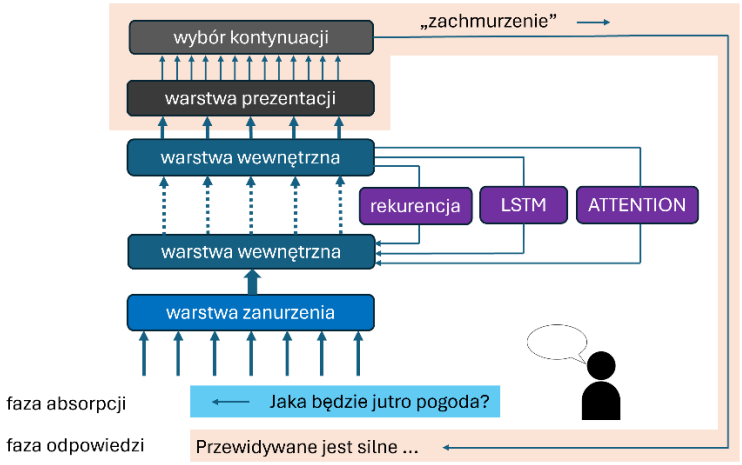
Tak więc sieci neuronowe stosowane w modelach języka nie są sieciami klasycznymi. Jedno z architekturalnych rozszerzeń klasycznej sieci neuronowej, tzw. sieć rekurencyjna, pozwala na przyrostowe budowanie i rozwiązywanie problemu w postaci kroków w ten sposób, że wyjście poprzedniego kroku stanowi część wejścia dla kroku następnego. Podstawowa wersja sieci rekurencyjnej posiada tę nieprzyjemną własność, że w miarę wykonywania kolejnych kroków gubi informację (zatraca wiedzę) o wynikach kroków poprzednich w tempie wprost proporcjonalnym do ich odległości od kroku bieżącego. Krótko mówiąc, im wcześniej wykonany został dany krok, tym mniejszy będzie jego wpływ na przyszłe kroki, przy czym zjawisko to nie zależy ani od struktury wejścia ani od wyników poszczególnych kroków, lecz bierze się z samej natury sieci. Niezależnie od architektury sieci (rekurencja nie jest tu ostatnim rozwiązaniem), iteracyjne wprowadzanie danych naturalnie wymusza preferencję dla danych świeżych, o ile nie zastosujemy specjalnych mechanizmów kompensujących tę tendencję. Zauważmy, że w wielu aplikacjach jest ono naturalne i pożądane. W modelowaniu języka bywa jednak tak, że istotność słów w zdaniu, zdań w wypowiedzi czy akapitów w dłuższej historii nie zawsze stanowi prostą funkcję chronologii. Czasem wracamy do tematu, przywołujemy dyskusję ze wstępu lub zapowiadamy coś, co zostanie wyjaśnione później. Można powiedzieć więcej: w najbardziej interesujących przypadkach rozumienia języka (i konwersacji) umiejętność skupienia się na istocie rzeczy, niezależnie od miejsca jej naświetlenia w chronologii konwersacji, stanowi największe wyzwanie.

Dwa kolejne uzupełnienia/rozszerzenia sieci neuronowych, bez których przełom w używaniu ich do modelowania języka nie byłby możliwy, noszą nazwę długiej pamięci krótkoterminowej (LSTM) oraz

[43] Jest to pierwsze miejsce, gdzie mówimy o semantyce w kontekście sieci neuronowych, warto więc zdać sobie sprawę z faktu, że chodzi tu (jedynie) o syntaktyczną reprezentację czegoś, co odpowiada ludzkiej interpretacji znaczeń. Słowa „perfumy" oraz „aromat" znajdują się blisko siebie w przestrzeni zanurzenia dla sieci neuronowej, gdyż wynika to z naszego przydzielenia znaczeń tym napisom. Gdyby pewnego dnia w korpusie dostępnym sieci konsekwentnie wymienić słowo „aromat" na „fetor", sieć nie zauważyłaby zmiany. Wiedza ta, czysto syntaktyczna, da się oczywiście wydobyć ze statystyk korpusu, tak więc procedury zanurzania podatne są na techniki trenowania sieci neuronowych. Problem jest traktowany osobno, gdyż można go rozwiązać wstępnie, tak że reszta procesu trenowania modelu języka przebiega niezależnie.

skupienia uwagi (ATTENTION).[44] W pierwszym przypadku chodzi o mechanizm, który umożliwiłby selektywne pamiętanie wyników potencjalnie zamierzchłych poprzednich kroków, tak by dało się zapewnić ich wpływ na kroki bieżące, pomimo przyrodzonej tendencji sieci do zapominania starych obliczeń. Drugi mechanizm ma na celu skupianie się na istotnych fragmentach przetwarzanej sekwencji pozwalając wzmacniać ich wpływ na wynik względem innych fragmentów. Okazał się on na tyle silny, że pozwolił zrezygnować z rekurencyjnej organizacji sieci, co drastycznie przyspieszyło uczenie modelu przez umożliwienie równoległego trenowania różnych fragmentów sieci. Typ sieci neuronowej oparty na takiej organizacji nosi nazwę transformera i leży u podstaw współczesnych modeli języka, jak ChatGPT.

Techniczne aspekty rzeczonych rozszerzeń są raczej zawiłe i techniczne, nie będziemy zatem omawiać ich szczegółowo. Wspomnę tylko, że ich forma, podobnie jak forma wszelkich możliwych użytecznych modyfikacji klasycznej koncepcji sieci neuronowej, ograniczona jest pewnym standardowym wymaganiem tamującym kreatywność w zakresie rewolucjonizowania tej koncepcji. Chodzi o to, że cokolwiek nowego nie wymyślimy, sieć wyposażona we te wszystkie nowalijki musi pozwalać się uczyć! Oznacza to tyle, że wszelkie parametry, które po zakończeniu procesu nauczania reprezentować będą wiedzę sieci muszą posiadać formę wag, których wpływ na zachowanie sieci podlega pewnym ścisłym matematycznym prawom określającym realizowalność i skuteczność algorytmu uczenia, czyli algorytmu wstecznej propagacji. Dotyczy to również dodatkowych parametrów kontrolujących długą pamięć krótkoterminową oraz skupienie uwagi.



Rysunek 14. Uproszczony schemat sieci neuronowej implementującej konwersacyjny model języka.

Przypomnijmy. Proces trenowania sieci neuronowej polega na dostarczaniu jej przykładów danych wejściowych oraz poprawnych (oczekiwanych) wyników klasyfikacji odpowiadających tym przykładom. W przypadku generatywnego modelu języka, korpus dostarcza przykładów tekstów oraz statystyk ich kontynuacji, które służą jako wzorce wartościowań. Ważnym rekwizytem jest tu funkcja błędu pozwalająca sieci wymiernie (numerycznie) oszacować rozmiar odstępstwa od oczekiwanego wyniku. Wartość tej funkcji używana jest przez algorytm wstecznej propagacji, który przebiega, jak wskazuje nazwa, wstecz – od wyniku do wejścia – mozolnie korygując wagi (parametry) napotkane po drodze. Warunkiem jego skuteczności jest dość silne ograniczenie sposobu, w jaki parametr może wpływać na funkcję kontrolowanego przezeń elementu sieci. W standardowym przypadku elementem tym jest neuron a parametrem waga, lecz nie musi tak być w przypadku rozszerzeń, które wprowadzają dodatkowe elementy kontrolowane parametrami podobnymi do wag neuronów. Trenowanie modeli języka jest kosztowne

[44] Angielskojęzyczne terminy to „Long Short-Term Memory (LSTM)” oraz „Attention” (czyli „Uwaga”). W tym drugim przypadku chodzi (mówiąc nieformalnie) o skupienie uwagi na istotnym fragmencie tekstu. Magnus Ekman, Learning Deep Learning. Addison-Wesley, 2022.

ze względu na astronomiczną liczbę wag (z których teoretycznie każda powinna być uaktualniona po każdym kroku uczenia) a także olbrzymią liczbę kroków (wymaganych dla precyzyjnego dostrojenia wag), przy czym należy pamiętać, że każdy krok wymaga przepuszczenia przez sieć olbrzymiej liczby przykładów z korpusu. Jest to proces niebywale czasochłonny, o ile nie posiadamy silnych narzędzi umożliwiających efektywną reprezentację sieci w sprzęcie i jej masywne równoległe przetwarzanie.

Rysunek 14 pokazuje uproszczony schemat sieci neuronowej realizującej konwersacyjny model języka. W fazie absorbcji sieć przyjmuje wypowiedź interlokutora, która wprowadzana jest cyklicznie w wejścia, przy czym wczytanie pojedynczego słowa traktowane jest jako pełny krok sieci powodujący przejście sygnałów przez wszystkie wewnętrzne warstwy. Na tym etapie sieć nie generuje sygnałów wyjściowych ustawiając jedynie wagi (potencjometry) opisujące stan LSTM i ATTENTION, co wprowadza model w kontekst konwersacji. W fazie odpowiedzi, model generuje kolejne słowa według zasady maksymalizacji prawdopodobieństw. Każde nowe słowo powiększa rozmiar frazy wejściowej dla wygenerowania następnego słowa, zgodnie z zasadą generatywnego modelu języka, którą opisaliśmy wcześniej.

Niezależnie od liczby parametrów, które można uznać za przechowalnię wiedzy wyuczonej przez neuronowy model języka, istotny jest rozmiar korpusu, na którym rzeczony model się uczy. Dla GPT wersji 3, rozmiar ten szacuje się na 45 terabajtów (45,000,000,000,000 znaków) tekstu pochodzącego z różnych źródeł (dostępnych w internetowych bazach danych), jak na przykład Wikipedia oraz książki.[45] Należy uzmysłowić sobie zakres informacji statystycznej dostępnej w takim korpusie. „Solaris" Stanisława Lema[46] obejmuje około 350 tysięcy bajtów, zatem rozmiar korpusu pokrywa około 130 milionów takich książek. Pojedynczy człowiek, żyjąc 90 lat i czytając jedną książkę dziennie, byłby w stanie przeczytać ich nieco ponad 30 tysięcy, czyli około 0,02% korpusu.

Niezależnie od sposobu, w jaki model posługuje się korpusem, jedno nie ulega wątpliwości: rzeczony korpus zawiera olbrzymią część dorobku naszej cywilizacji zakodowaną w postaci symboli: liter, słów, znaków. Symbole te są z sobą skorelowane na niezliczone sposoby i gdzieś w tej astronomicznej gmatwaninie statystycznych zależności zawiera się z grubsza wszystko, co kiedykolwiek, z odrobiną sensu, wypowiedział człowiek. Nie mając bladego pojęcia o znaczeniu tej informacji, nie mając bladego pojęcia o czymkolwiek i nie będąc w stanie pojąć, co to znaczy mieć blade pojęcie o czymkolwiek, model karmi się korpusem (gra słów niezamierzona) tworząc misterne, wewnętrzne i zasadniczo niedostępne nam schematy, według których ciąg pewnych napisów wiąże się z innym ciągiem, a z tego związku wynika, że w reakcji na pewien ciąg, należy wygenerować pewien inny ciąg, i tak dalej, czyniąc to z szybkością, której nie jesteśmy sobie w stanie wyobrazić. Wszystko sprowadza się do wymykającego się wszelkim analogiom tańca liczb choreografowanego bilionami wirtualnych potencjometrów misternie wykalibrowanych w procesie uczenia. Tak naturalista chciałby widzieć ludzki mózg – jako cyfrowy mechanizm, w którym jeden impuls powoduje drugi, drugi powoduje trzeci, a całość prowadzi do świadomości, myślenia, inteligencji, emocji, radości, strachu, zachwytu, cierpienia, kultury, czyli tego wszystkiego, co leży u źródeł obecnego sukcesu sztucznej inteligencji, mianowicie semantycznej treści zaklętej w martwą składnię karmiącego ją korpusu. Od momentu powstania pierwszych systemów perceptronowych ich twórcy chcieli tam widzieć neurony, najlepiej funkcjonalnie identyczne z tymi, z których zbudowane są nasze mózgi.

[45] Jako punktu odniesienia używamy korpusu/modelu angielskojęzycznego, gdyż najlepiej określa on zakres wysiłków w tej dziedzinie.

[46] Stanisław Lem, Solaris. Wydawnictwo Literackie, 2012.

Czy sztuczna inteligencja myśli?

Współczesne generatywne modele języka potrafią bez trudu zdać test Turinga. Jordan Peterson w rozmowie z Brianem Roemmele[47] porównał swoją konwersację z ChatGPT dotyczącą tematów z zakresu psychologii do dyskusji z ponadprzeciętnym studentem. Jest to bez wątpienia wielki sukces sztucznej inteligencji ziszczający wizje autorów i filmowców fantastyki naukowej, sporej części naukowców oraz wielu zwykłych ludzi. Czy wynika z tego, że komputery zaczęły w końcu myśleć?

Jeśli zdefiniujemy (ludzką) inteligencję jako cechę przejawianą przez każdy mechanizm czy ustrój, który zaliczy test Turinga, wówczas z definicji zmuszeni będziemy przyjąć, że ChatGPT ją posiada. Jeśli zdefiniujemy myślenie jako działalność mechanizmu czy ustroju owocującą przejawieniem ludzkiej inteligencji, wówczas z definicji zmuszeni będziemy przyjąć, że ChatGPT myśli. Natura definicji jest taka, że opisują one dokładnie to, co wyrażone jest w ich treści. Nie wnoszą nic nowego oferując jedynie skróty. Próbując opisać istotę naszych ludzkich poczynań przy pomocy tych własności mechanizmów, które uda nam się zdefiniować, nie dowiemy się niczego o tym, czego nie potrafimy zdefiniować jako własność mechanizmu. Alan Turing był matematykiem i rozumiał doskonale co to definicja. Wbrew późniejszym ewangelistom celebrującym jego test konstatował jedynie pewną tautologię orzekającą, że jeśli kiedyś przypadkiem komputer zacznie nagle gawędzić niczym człowiek, to wypadnie uznać, że faktycznie, komputer potrafi gawędzić niczym człowiek. Podobnie jak wielu współczesnych mu badaczy matematycznej teorii obliczalności, Turing podejrzewał, że absolutnie każdy proces efektywnie realizowalny w świecie da się opisać przy pomocy komputerowego programu, którego wykonanie będzie równoważne realizacji rzeczonego procesu.[48] Tak to już jest, że gdy otrzymamy w prezencie nowy lśniący młotek, odczuwamy niepohamowaną ochotę stuknąć nim w każdy przedmiot, jaki pojawi się w zasięgu naszej ręki. Teza ta dotyczyła także wszystkich funkcji ludzkiego mózgu, gdyż założenie naturalizmu i fizykalizmu myślenia i świadomości było w owych latach nawet bardziej oczywiste niż dziś, co czyniło tautologię Turinga rodzajem refleksji nad prozaiczną obliczalnością naszych namiętności. W dzisiejszych czasach takie założenie stanowi przesłankę spekulacji, że nasz świat jest komputerową symulacją (grą) zabawiającą znudzonego nastolatka wysoko rozwiniętej cywilizacji z innego uniwersum.[49]

Pomimo braku rzetelnego modelu procesu myślenia, który dałby się przyłożyć do schematu sieci neuronowej, wiemy na pewno, że nasz mózg funkcjonuje inaczej niż sieć neuronowa czy model języka. Wiemy o tym choćby dlatego, że mózg nie jest w stanie przetworzyć korpusu tekstów o zawartości wykraczającej poza kilka stronic i zapamiętać na wieki cyfrowego ekstraktu z wszystkiego, co w nim ważne. Nie posiadamy żadnego modelu ludzkiej świadomości, nawet takiego, który przyśniłby się kilku filozofom na raz w podobny sposób. Nie istnieje sposób na zdiagnozowanie świadomości w maszynie, choćby poszlakowy, jak tautologia Turinga, gdyż zupełnie nie wiadomo, co mianowicie należałoby zdefiniować, by stworzyć iluzję testu na świadomość. O ile bowiem (pewne) objawy inteligencji dadzą się postrzec przez postronnego obserwatora, o tyle świadomość stanowi najbardziej intymny, wewnętrzny i całkowicie bezobjawowy atrybut naszego istnienia.

Współczesne filozoficzne teorie na temat miejsca, gdzie świadomości należy szukać różnią się drastycznie. Nie sposób naszkicować ich tu choćby pobieżnie i każda próba skrótu doprowadzi do groteski w

[47] Jordan Peterson interviewing Brian Roemmele, ChatGPT and the Dawn of Computerized Hyper-Inteligence, YouTube, https://www.youtube.com/watch?v=S_E4t7tWHUY .
[48] B. Jack Copeland et al., Computability: Turing, Gödel, Church, and Beyond, MIT Press, 2013.
[49] Rizwan Virk, The Simulation Hypothesis: An MIT Computer Scientist Shows why AI, Quantum Physics and Eastern Mystics All Agree We Are in a Video Game, Bayview Books, 2019.

stylu filozofii ruchliwych kartofli.[50] Nawet najbardziej radykalne idee, np. iluzjonizm[51] upierający się, że świadomość nie istnieje jako fenomen, lecz wyłącznie jako złudzenie, dostrzegają pewien problem z należytym wyjaśnieniem cóż to takiego owo złudzenie.

Brak definicji świadomości (i brak możliwości stworzenia takiej definicji na użytek jej zewnętrznego obserwatora) powoduje, że nie sposób potraktować zagadnienia w sposób uczciwie naukowy. Nauka dotyczy bowiem zjawisk i fenomenów zachodzących i istniejących obiektywnie. Odbywa się to tak, że najpierw postrzegamy coś subiektywnie, na przykład magnes zawieszony na sznurku ustawia się w określonej pozycji. Potem tworzymy hipotezy i próbujemy je weryfikować przy pomocy eksperymentów. Na pewnym etapie postulujemy obiektywną teorię: Ziemia posiada globalne pole magnetyczne. Pokazujemy wszem i wobec jak można zawiesić magnes na sznurku i sprawdzić, że to co mówimy ma sens. Znając geometrię naszej planety potrafimy przewidzieć, jak magnes ustawi się w dowolnym jej miejscu. W przypadku świadomości, jej subiektywne postrzeganie, nasza introspekcja, to już wszystko. Nic więcej nie potrafimy z tego wyprowadzić. Na dobrą sprawę nikt nie jest pewien, czy nie jest przypadkiem jedyną świadomą istotą we wszechświecie. Solipsyzmu nie da się logicznie zanegować i jedyną obroną przed nim jest wiara wynikająca z obserwacji, że osobniki, z którymi obcuję twierdzą, że także są świadome i wyglądają, na pierwszy rzut oka, podobnie jak ja.

Z braku możliwości dokonania autentycznego naukowego postępu musimy się zadowolić nieskażoną metafizyką.[52] Oto co pisze jeden z najbardziej prominentnych współczesnych filozofów od świadomości, Thomas Nagel:

> „Organizm posiada świadomość, jeśli istnieje jakiś fenomen odpowiadający poczuciu bycia tym organizmem, odczuwaniu bycia sobą jako rzeczony organizm".[53]

Zgadza się – posiadam coś takiego, a przynajmniej tak mi się wydaje. Nie wynika z tego jednak definicja, która pozwoliłaby obiektywnie ustalić (albo choćby przyjąć dla świętego spokoju – jak w przypadku testu Turinga), że inteligentnie rozmawiająca ze mną maszyna jest także świadoma. Rzekomo świadoma maszyna, jak każde z nas, musi uczynić to sama i może mi co najwyżej o tym powiedzieć. No i w każdym przypadku, podobnie jak dowolny inny interlokutor, maszyna może mi powiedzieć to, co jej przyniesie na (elektroniczny w tym przypadku) język jej (umowna w tym przypadku) ślina.

David Chalmers, uważany za wiodącego eksperta od metafizyki świadomości,[54] dzieli poszukiwania między dwa terytoria: łatwe i trudne. Problemy łatwe w badaniu świadomości to takie, nad którymi w ogóle potrafimy się pochylić i próbować je rozwiązywać przy pomocy arsenału narzędzi naturalistycznych, którymi posługują się nauki przyrodnicze. Nie muszą być faktycznie łatwe, ale wiadomo o nich tyle, że jeśli spędzimy dostatecznie wiele czasu i środków, to osiągniemy tam postęp. Przykładowo, należą do nich zjawiska związane z percepcją i przetwarzaniem sygnałów odbieranych przez nasze zmysły. Wolno nam na przykład powiedzieć, że jesteśmy świadomi sygnału odbieranego przez nasz nos, w podobny sposób, jak czujnik dymu jest świadomy, że coś się pali. Stwierdzenie takie oznacza w swoim kontekście, że jakiś namacalny fizyczny proces wykrył zmianę chemii w otoczeniu, co z kolei powoduje wygenerowanie sygnału wyzwalającego jakąś akcję. Do problemów trudnych, którymi nie potrafimy

[50] Stanisław Lem, Dzienniki Gwiazdowe (Podróż Dwudziesta Piąta), Wydawnictwo Literackie, 2012.
[51] Keith Frankish, Illusionism as a Theory of Consciousness, Journal of Consciousness Studies, 23, 11-12, pp. 11-39, 2016.
[52] David Chalmers, David Manley, and Ryan Wasserman, Metametaphysics, Oxford University Press, 2009.
[53] Thomas Nagel, Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False, Oxford University Press, 2012 (moje tłumaczenie).
[54] David Chalmers, The Conscious Mind: In Search of a Fundamental Theory, Oxford University Press, 1996.

się zająć poza sferą metafizycznych spekulacji, zalicza Chalmers wyjaśnienie fenomenów introspekcji, poczucia własnego istnienia, odczuć, wrażeń, czyli autentycznej świadomości.

Przed chwilą, po skończeniu powyższego akapitu, dopiłem kawę i wyszedłem przed dom. Jest słoneczny poranek, śpiewa kos, po torach przejeżdża Pendolino. Widzę przed sobą zielone pole i czuję jego majowy zapach po wczorajszym koszeniu. Po prawej krubiński lasek, a na wprost, za torami, niepodrabialny, na swój jedyny sposób płaski, krajobraz Mazowsza rozciągający się po horyzont, z odległą, ukrytą nitką drogi do Płońska, po której sporadycznie snują się pojazdy. Wspominam brata, dla którego to, co wiąże materię naszego świata z ułudą ludzkiego istnienia, jak chcieliby orędownicy iluzjonizmu, zakończyło się dwa lata temu na tych właśnie torach. Myślę o wszystkim naraz, jak zwykle po kawie i – jak każdy człowiek, włączając skonfundowanych filozofów – chciałbym wiedzieć, dlaczego właściwie myślę i co to wszystko znaczy. Jedyna pociecha w mojej irytującej niewiedzy bierze się stąd, że wypadkowa intelektualnych fajerwerków najwybitniejszych myślicieli wszystkich epok wynosi w tej materii niemal dokładnie zero.

Pod koniec moich informatycznych studiów, w połowie lat siedemdziesiątych, byłem przekonany, że procesy przebiegające w mózgu stanowią rodzaj obliczenia realizowalnego przez komputer i wierzyłem, że jeśli jeszcze trochę pogmatwam moje programy to zbliżą się one do myślenia i pozyskają świadomość na zasadzie czystej komplikacji. Coś takiego nazywa się emergencją.[55] Filozofia, a nawet okazjonalnie fizyka, posługuje się tym terminem kiedykolwiek musi wytłumaczyć coś, co stanowi całkowicie nową jakość w skądinąd znanej i zrozumiałej strukturze, której to nowej jakości nie można wyprowadzić przez stopniowe dekonstruowanie i analizowanie rzeczonej struktury od podstaw, przy pomocy elementarnych, powiązanych ze sobą i zrozumiałych kroków. Weźmy dla przykładu samochód. Gdy pokażemy gotowy pojazd komuś, kto go nigdy przedtem nie widział, dajmy na to przybyszowi z egzotycznej planety, i zademonstrujemy jego funkcje, delikwent ma prawo zapytać skąd to się wzięło. Gdy następnie pokażemy mu wszystkie niezbędne, elementarne, materialne rekwizyty w postaci rud metali, ropy naftowej, itd., które leżą u absolutnego źródła procesu technologicznego prowadzącego do końcowego produktu bez wytłumaczenia szczegółów, nic nie wyjaśnimy i co najwyżej zainicjujemy łańcuch spekulacji w głowie przybysza (zakładając, że posiada on głowę). Na skutek nieporozumienia, może on dojść do wniosku, że jeśli zbierzemy w jedno miejsce te wszystkie materiały i wrzucimy je do jednego wielkiego pojemnika, a potem nim dla pewności potrząśniemy, to nastąpi emergencja, czyli pojazd zmaterializuje się sam, gdyż istnieje gdzieś prawo, które mówi, że określona konfiguracja materii samoczynnie prowadzi do powstania nowej jakości w postaci gotowego i w pełni funkcjonalnego samochodu.

Naturalistyczne objaśnienia genezy świadomości przebiegają z grubsza według tej samej linii. Jak wiemy, wszystko przecież dzieje się „samo" zgodnie z zasadami neodarwinizmu. Różne kręgi naukowców przejawiają różny stopień oporu przed tą doktryną, często tłumionego instynktem

[55] Mała dygresja. Nie posługując się językiem polskim w moich naukowych studiach i opracowaniach (od czasu emigracji w roku 1984), nie zdawałem sobie sprawy z istnienia takiego słowa. Przez pewien czas poszukiwałem polskiego odpowiednika angielskiego „emergence", lecz oficjalne słowniki nie okazały się pomocne sugerując znaczenia jak „powstawanie", „wystąpienie", itd., które nie oddają konotacji, jakie niesie angielski termin w filozofii i fizyce. Wymyśliłem więc własne słowo „samopowstanie", po czym coś mnie tknęło i udałem się do Wikipedii pod hasło „Emergence", a następnie poprosiłem o polskojęzyczną stronę na ten sam temat. No i wszystko się wyjaśniło. Nawet sprawdzarka pisowni w moim Wordzie oblizała się z zachwytem, choć krzywi się na niektóre słowa, które mnie z kolei wydają się całkowicie legalne. Wysłuchując niedawno pogadanki Profesora Bralczyka, mojego starszego kolegi z liceum, dowiedziałem się, że musimy się godzić z podobnymi wtrętami, jeśli nic innego nie da się zrobić. Przyznam jednak, że „samopowstanie" podoba mi się bardziej niż „emergencja" (o której nie uczyliśmy się z profesorem Bralczykiem w liceum). Być może jest zbyt blisko „objawienia" lub „zmartwychwstania", co może razić uszy niektórych filozofów, podczas gdy „emergencja" brzmi tajemniczo i naukowo. No ale tak naprawdę chodzi tu o coś bardziej przypominającego cud niż naturalny, fizykalnie wytłumaczalny mechanizm.

samozachowawczym, co zdaje się ją umacniać w sposób przekornie cykliczny. Istnieją tym niemniej poważne obiekcje przeciwko forsowaniu emergencji jako metodologii konstruowania naukowych objaśnień zjawisk, o których nie posiadamy bladego, czy choćby jako tako satysfakcjonującego, pojęcia. Niektóre z tych obiekcji pochodzą nawet ze środowisk biologów, gdzie często prowadzą do ostracyzmu i stygmatyzowania, szczególnie gdy osoby je głoszące nie kryją się z poglądami o charakterze religijnym – nawet jeśli ich argumentacja jest pozbawiona demagogii.[56]

Krytykowanie naturalizmu i fizykalizmu przez filozofów przychodzi lżej, gdyż filozofowi łatwiej zanegować popularny pogląd naukowy bez konieczności dostarczania alternatywy w postaci gotowej i precyzyjnej teorii. Bogiem a prawdą, status quo na temat szczegółów funkcjonowania świata, czy nawet ich ramifikacji, nie istnieje wśród współczesnych fizyków i jedynym miejscem, gdzie daje się dostrzec instynkt stada z odruchową obroną pozycji jest właśnie neodarwinizm w biologii. Według słów Davida Berlinskiego,[57] jest to coś w rodzaju „linii partii", czyli oficjalnego stanowiska biura politycznego, w które członkowie utracili wiarę i którego nie muszą przestrzegać w prywatnej konwersacji, choć nie powinni się wychylać w publicznych miejscach z niezależnymi, kontrowersyjnymi opiniami.

Tak więc niewielu filozofów i bynajmniej nie wszyscy fizycy, szczególnie pośród tych, którzy spoglądali w stronę trudnych problemów świadomości, upiera się, że współczesna fizyka jest w stanie wyświetlić te problemy. Jeden z kierunków w filozofii świadomości, misterianizm, którego prominentnym reprezentantem jest Colin McGinn, głosi, że ludzki umysł nie jest wyposażony w środki, które umożliwiłyby nam rozwiązanie problemu świadomości.[58] Inny reprezentant tego samego ogólnego kierunku, nie żyjący już Jerry Fodor, w swoich próbach sformułowania algorytmicznych zasad funkcjonowania umysłu doszedł do wniosku, że procesy opisujące świadomość nie mogą być redukowalne do funkcji neuronów i synaps.[59] W przypadku Chalmersa, konkluzja jest taka, że teorie fizyki nie mogą być teoriami wszystkiego i że do zrozumienia świadomości potrzebny jest dodatkowy fundamentalny składnik, o którym nie mamy zielonego pojęcia. Składnik ten, według Chalmersa, nie jest redukowalny do niczego, co mogłoby nam dzisiaj przyjść do głowy. Wypada zauważyć, że przy swoim autorytatywnym wsparciu dla tezy, która ma prawo łatwo skojarzyć się ze Stwórcą, Chalmers jest tak odległy od spirytualizmu, jak tylko można to sobie wyobrazić. Będąc filozoficznym dyletantem, zakwalifikuję go po prostacku jako dualistę, czyli zwolennika separacji ducha i materii (jak Kartezjusz).[60] Jest to klasyfikacja mało precyzyjna, gdyż wariantów dualizmu w filozofii istnieje kilkadziesiąt.

Jak argumentuje Wolfgang Smith, matematyk i filozof fizyki,[61] kartezjański dualizm w istotnym zakresie pozostaje fundamentalnym rekwizytem współczesnej nauki pomimo jej materialistycznego (czy naturalistycznego) credo, którego litera ulega w ostatnich czasach rozmyciu. Odseparowanie rzeczy materialnych (*res extensae*) od sfery umysłu (*res cogitantes*) jest jego zdaniem dokładnie tym, co umożliwiło zastosowanie ścisłych reguł (czyli matematyki) do opisu (naturalistycznej części) świata i doprowadziło do sukcesu w tej dziedzinie ogłaszanego przez naturalistów jako spektakularny. Mechanika kwantowa, uzależniając całą fizykę od roli obserwatora, zasadniczo opiera się na tym podziale. Smith nie zgadza się z kartezjańskim dualizmem, gdzie *res cogitantes* pojawiają się wyłącznie w umyśle, podczas gdy *res extensae* tworzą cały (zewnętrzny) świat. Jego idea wertykalnej przyczynowości osadza *res cogitantes* w rzeczywistym świecie, który składa się z dwóch części: korpor(e)alnej, w której funkcjonujemy my razem z naszymi zmysłami i percepcją, oraz fizycznej, w której funkcjonują prawa fizyki. Udało nam się

[56] Michael J. Behe, Darwin Devolves, HarperOne, 2019.
[57] David Berlinski, The Devil's Delusion: Atheism and its Scientific Pretentions, Read How You Want, 2010.
[58] Collin McGinn, The Problem of Consciousness, Blackwell Pub, 1991.
[59] Jerry A. Fodor, The Modularity of Mind, MIT Press, 1983.
[60] Richard A. Watson, What Moves the Mind: An Excursion in Cartesian Dualism, American Philosophical Quarterly, University of Illinois Press, 19 (1), pp. 73-81, 1982.
[61] Wolfgang Smith, Physics and Vertical Causation: The End of Quantum Reality, Angelico Press, 2019.

dość gruntownie zbadać fizyczną część świata, gdyż po prostu okazała się łatwa, nie tykając (a nawet nie zauważając) tej drugiej części. Stąd między innymi wzięły się problemy z interpretacją mechaniki kwantowej. Po prostu poznaliśmy niechcący odrobinę więcej fizyki niż potrafiliśmy przełknąć upierając się, że nic więcej nie ma. Spektakularność naszego sukcesu w matematycznym opisywaniu świata jest więc względna. Po pierwsze, nie opisaliśmy wszystkiego, a po drugie, nie mamy bladego wyobrażenia w kwestii zakresu tego, czego opisać nam się dotąd nie udało.

Emergencja jako obrona przed niezrozumiałym stosowana bywa w sytuacji, gdy naukowa doktryna utrudnia pogodzenie się z brakiem środków na wnikliwe i wyczerpujące objaśnienie analizowanych zjawisk. Rozumowanie przebiega według następującego schematu: 1) dzieje się coś, czego nie rozumiemy, 2) nasz system wiedzy o świecie oparty jest na pewnych założeniach, które są niepodważalne, 3) nie potrafimy znaleźć objaśnienia obserwowalnego zjawiska przez jego elementarną analizę w ramach naszego systemu, 4) zatem zjawisko występuje na zasadzie emergencji, jako coś, co musi się pojawić samoczynnie. W ten sposób nasz system wiedzy zostaje rozszerzony o nowy aksjomat, który brzmi na przykład tak:

> „Świadomość pojawia się samoczynnie w dostatecznie skomplikowanym systemie."

Razem z nim przychodzi rekomendacja, by nie zaprzątać sobie głowy dalszym roztrząsaniem problemu.

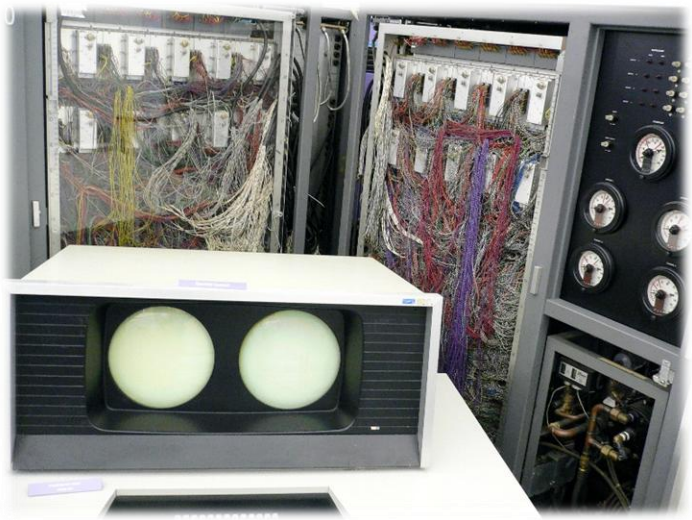
Nadużywanie pojęcia emergencji ułatwione jest jego częstym stosowaniem w sytuacjach, w których chodzi o zwykłe i naturalne pojawienie się pewnej funkcji, która w sposób całkowicie wytłumaczalny wynika z fizykalnej przyczynowości prostych zjawisk nie kryjąc żadnej tajemnicy. Powiedzenie, że Internet jest emergencją z równań Maxwella oznacza coś innego niż nonszalancka konkluzja, że ludzka świadomość pojawiła się jako emergencja z chemicznych własności węgla. W pierwszym przypadku posiadamy pełną wiedzę na temat każdego elementu łańcucha prowadzącego od dobrze znanych fizykalnych przesłanek do współczesnych technologii telekomunikacyjnych; w drugim, nie wiemy nawet, o czym mówimy. Ustawianie takich „obserwacji" w jednym szeregu to przejaw kultu terminologii, gdzie naukowemu nazwaniu zjawiska przypisuje się magiczną moc jego objaśnienia na zasadzie „naukowcy już tam wiedzą jak to się dzieje (albo dowiedzą się wkrótce) i nie tobie to kwestionować". Poszukując przykładów podobnych nadużyć natknąłem się na pewien artykuł, którego fragment (w moim tłumaczeniu, z moim podkreśleniem) pozwalam sobie przytoczyć poniżej.[62]

> „Gdy elektrony, atomy, jednostki lub społeczeństwa podlegają interakcjom między sobą lub ze środowiskiem, grupowe zachowanie całości różni się od zachowania jej części. Takie grupowe zachowanie nazwiemy emergentnym. Emergencja odnosi się zatem do zbiorowych zjawisk w złożonych systemach podlegających adaptacji, które to zjawiska nie występują w ich poszczególnych częściach. Przykłady zjawisk emergentnych pojawiają się wszędzie wokół nas: gromadzenie się ptaków, synchronizacja świetlików, kolonie mrówek, ławice ryb, tworzenie się dzielnic w miastach. Wszystko to dzieje się bez przywódców i bez centralnego sterowania. Przykładami emergencji są także: Wielki Wybuch, formowanie się galaktyk, gwiazd i planet, ewolucja życia na Ziemi, od jego początków do chwili obecnej, powstawanie protein, tworzenie się komórek, krystalizacja atomów w cieczy, nadprzewodnictwo elektronów w niektórych metalach, zmieniający się globalny klimat lub rozwój <u>świadomości</u> u niemowlęcia."

Rażąca jest niedorzeczność powyższej mieszanki, w której znajdujemy: 1) proste fizyczne zjawiska opisane precyzyjnymi teoriami: krystalizacja, nadprzewodnictwo; 2) zjawiska fizyczne, co do których istnieją częściowo spekulatywne teorie oparte na solidnych formalnych podstawach: formowanie się

[62] David Pines, Emergence: A Unifying Theme for 21st Century Science, Foundations & Frontiers of Complexity, Santa Fe Institute Bulletin, 28(2), 2014.

galaktyk, gwiazd, planet; 3) łatwo wytłumaczalne zjawiska biologiczne: gromadzenie się ptaków, synchronizacja świetlików, kolonie mrówek, ławice ryb; 4) cokolwiek mniej ogarnięte procesy biologiczne: ewolucja życia na ziemi, tworzenie się białek i komórek; 5) proste procesy społeczne nie sprawiające epistemologicznych kłopotów socjologom; 6) Wielki Wybuch, 7) rozwój świadomości u niemowlęcia; 8) zmieniający się globalny klimat – jakżeby inaczej? Konstrukcja tej absurdalnej papki ma na celu stworzenie wrażenia, że wszelkie zjawiska na świecie posiadają jedno doskonałe i w domyśle proste wytłumaczenie, mianowicie emergencję, i jedynym celem nauki w 21 wieku jest studiowanie tego unifikującego mechanizmu, co pozwoli nam rozwiązać wszystkie problemy filozoficzne i naukowe ratując przy okazji Planetę. Powstanie świadomości (u niemowlęcia) ma tu wynikać jako logiczna konsekwencja obserwacji, że o ile pojedynczy ptak gromadzić się nie potrafi, to pięć już tak.



Rysunek 15. CDC 6000 konsola oraz jednostka centralna z panelem chłodzenia (Creative Commons, Steve Jurvetson, Menlo Park, CA).

Z drugiej strony, spoglądając na współczesny komputer i próbując ogarnąć jego budowę oraz złożoność wykonywanych przezeń programów, trudno się oprzeć wrażeniu, że ta komplikacja może prowadzić do czegoś całkiem nowego, co pojawi się samo, na zasadzie emergencji, wynikając wyłącznie z komplikacji. Wrażenie się nasila, gdy pozwolimy się ponieść uduchowionemu entuzjazmowi filozofów „New Age" zilustrowanym powyższym cytatem i skonstatujemy, że nasz komputer połączony jest w światową sieć o rozmiarach i szybkości przesyłania informacji, jakie jeszcze kilka lat temu nie śniły się filozofom. Nie śniły się także naukowym fantastom sprzed siedemdziesięciu lat, którzy już wtedy snuli wizje komputerowego buntu wywodząc, że odpowiednio duży konglomerat kabli i tranzystorów (a dokładniej lamp) będzie się rządził swoimi własnymi prawami odmiennymi od przyświecających elektronice i telekomunikacji równań Maxwella.

Utkwił mi w pamięci moment, gdy jako młody i wystraszony student oczytany w książkach Lema oraz kilku pomniejszych futurystów znalazłem się w pomieszczeniu najpotężniejszego komputera w Polsce, którym był wtedy Cyber CDC 6000 (Rysunek 15) zainstalowany w Instytucie Badań Jądrowych w Świerku, gdzie kilka dni później zostałem przyjęty do elitarnej grupki tzw. analityków systemowych. Było to niewątpliwie jedno z najdonioślejszych wydarzeń w moim życiu (zapamiętałem je wyraźniej niż o wiele późniejszą obronę pracy doktorskiej), gdyż wydawało mi się wtedy, że oglądam najważniejsze z osiągnięć naszej cywilizacji. Gdy pokazano mi plątaninę drutów pod pokrywą centralnego procesora (o rozmiarach wielkiej szafy przypominającej swym surowym kształtem monolit z „Odysei Kosmicznej") straciłem oddech. Nie miałem wątpliwości, że to urządzenie potrafi dokonywać cudów i że jeśli nie ja sam, to z pewnością ktoś niebawem nauczy je myśleć.

W porównaniu do mocy obliczeniowej laptopa, na którym piszę te słowa, tamten superkomputer był mniej wydajny niż dziecinna hulajnoga w konfrontacji z promem kosmicznym. Mój laptop jest około 10 tysięcy razy szybszy, posiada 40 tysięcy razy więcej pamięci operacyjnej i 50 tysięcy razy więcej pamięci dyskowej, zatem nawet takie porównanie nie odzwierciedla różnicy. Całość oprogramowania tamtego systemu, czyli system operacyjny, kompilatory i wszelkie standardowe aplikacje, zajmowała mniej miejsca niż trzy fotografie mojego psa, które kątem oka dostrzegam na pulpicie.

Spróbujmy zrozumieć drogę przebytą od tamtych czasów. Jasne, że mamy teraz coś nowego. Komputery stały się nieporównanie bardziej wydajne, choć, mówiąc szczerze, zerkając na powyższe liczby, oczekiwałbym bardziej spektakularnych różnic w funkcjonalności. W tamtych czasach programowało się inaczej i mały skrawek pamięci w zupełności wystarczał do wielu praktycznych celów. Pomimo że dobrze rozumiem jak to wszystko funkcjonuje, ciągle nie mogę się pogodzić z faktem, że mój laptop wymaga 16 gigabajtów pamięci operacyjnej, by bez frustrującego oczekiwania realizować kilka w sumie banalnych funkcji.

Tak więc mamy teraz te wszystkie wspaniałe aplikacje, gry, grafikę, a przede wszystkim Internet, który zrewolucjonizował większość aspektów naszego życia, czasem na lepsze, czasem na gorsze – zależnie od wieku i gustu. Zmieniło się tak wiele, że nie wiadomo, od czego zacząć, więc darujmy sobie historię, którą z grubsza wszyscy znamy. Cokolwiek się nie zmieniło, wszystko to da się zrozumieć i wytłumaczyć krok po kroku, poczynając od pierwszego urządzenia, które zasługiwało na miano komputera po dzisiejsze tensorowe procesory wspierające generatywne modele języka. Nie wystąpiła po drodze żadna emergencja. W żadnym z kroków postępu komputer nie zaczął nagle myśleć i nie pojawiła się u niego świadomość. Systemy sztucznej inteligencji, włączając modele języka, też rozwijały się stopniowo i te pierwsze z nich, jak wspomniana wcześniej Eliza, nie budziły kontrowersji ani zaniepokojenia swoim potencjałem intelektualnego konkurowania z ludźmi.

Idea testu na komputerową inteligencję nie była jedynym wkładem Alana Turinga w teoretyczną informatykę. Nieporównanie ważniejszą częścią jego dorobku jest formalizacja pojęcia obliczenia w postaci tak zwanej maszyny Turinga, która dostarcza teoretykom informatyki potężnego matematycznego narzędzia dla szacowania formalnych możliwości absolutnie wszystkich komputerów funkcjonujących według ogólnej zasady von Neumanna. Dotyczy to w szczególności wszystkich cyfrowych komputerów, które potrafimy zbudować, włączając komputery kwantowe oraz wszelkie specjalizowane procesory, na których realizują się sieci neuronowe.

Jako użytkownicy komputerów, zdajemy sobie intuicyjnie sprawę z faktu, że wszystkie współczesne komputery są z grubsza równoważne pod względem zakresu dostępnych im funkcji. Zakupując nowy laptop nie interesujemy się zwykle repertuarem instrukcji maszynowych oferowanych przez procesor, gdyż wierzymy, że przykrywająca go warstwa oprogramowania spowoduje, że nasze aplikacje będą działać jak poprzednio, co najwyżej szybciej. Czasem wahamy się między Apple (macOS) i Microsoft (Windows), gdzie zarówno sprzęt jak i system operacyjny różnią się drastycznie, ufając, że funkcjonalnie, z punktu widzenia praktycznych możliwości, wybór nie ma istotnego znaczenia.

Jasne, że jeden komputer może być szybszy niż drugi, może posiadać mniej lub więcej pamięci, może zajmować mniej lub więcej miejsca w plecaku, pracować dłużej lub krócej bez doładowywania baterii, pełnić funkcje nie związane z jego podstawową rolą, na przykład podnosić status społeczny właściciela. Z punktu widzenia matematyka analizującego zakres możliwości komputera są to wszystko szczegóły dekoracyjne, które rzeczony matematyk pospiesznie pominie. Jego celem będzie ustawienie kreski oddzielającej problemy, które komputer potrafi rozwiązać od tych, które są dla niego niedostępne. Dokonując takiej analizy musimy najpierw ustalić jak reprezentować problemy oraz ich rozwiązania by reprezentacja była prosta i jednocześnie stanowiła faktyczną reprezentację wszystkich problemów, które ktoś kiedykolwiek chciałby komputerowi przedłożyć. Na pierwszy rzut oka wydaje się to niemożliwe,

gdyż mnogość sposobów, na które przedstawiamy komputerowi zadania i interpretujemy jego odpowiedzi zdaje się wykluczać ich jednorodną, prostą reprezentację.

Przypomnijmy sobie stare filmy fantastyczno-naukowe, na których operator komputera zainstalowanego na statku transgalaktycznym, przedstawionego jako zestaw metalowych szaf pokrytych migoczącymi bez sensu światełkami, przyciska guzik, wprowadza w otwór podziurkowaną kartę lub tasiemkę i otrzymuje w odpowiedzi inny podziurkowany kawałek papieru, na który spogląda okiem eksperta orzekając: „systemy zachowania życia funkcjonują w normie". Przez długi czas fantastom wydawało się, że w konwersacjach z komputerem zawsze skazani będziemy na posługiwanie się trywialnymi i uciążliwymi interfejsami, których interpretacja wymagać będzie tajemnej wiedzy operatora w białym fartuchu tłumaczącego komputerowe na ludzkie i odwrotnie. Nawet jeśli filmowy komputer okazjonalnie przemówił ludzkim głosem, nawet jeśli czasem wysłuchał czegoś, co kapitan statku miał akurat do powiedzenia, rekwizyt papierowych kartek i tasiemek pokutował długo jako jego charakterystyczny atrybut.

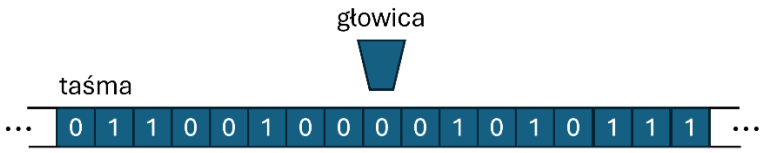
Tymczasem wygląda to tak, że o ile problemy związane z tworzeniem nowoczesnych interfejsów (grafika, głos, drukarki 3D) mogą zawierać poważne wyzwania technologiczne, są one jednocześnie całkowicie banalne z punktu widzenia matematyka. Wszystko sprowadza się bowiem do kodowania, czyli tworzenia mechanizmów zamiany pewnych sygnałów na ciągi bitów w pamięci maszyny i na odwrót. Sposoby (algorytmy) dokonywania takich konwersji mogą być ciekawe dla inżyniera, lecz jeśli już istnieją i działają, możemy przestać o nich myśleć. Nie wprowadzają one żadnej nowej jakości w fundamentalny problem intrygujący naszego matematyka, gdyż można je pominąć przyjmując, że każde zadanie przedstawione komputerowi do rozwiązania i każdy wynik wyprodukowany przez tenże komputer wygląda jak sekwencja bitów w pamięci. Uogólniając, powiemy, że każdy rodzaj kodowania (i dekodowania), o ile jest ono dobrze określone (znamy jego algorytm i potrafimy go zawsze zastosować) stanowi nieistotny etap w obliczeniach komputera, gdyż zmienia jedynie formę zapisu informacji bez dotykania jej treści.

Maksymalne uproszczenie interfejsu komputera (bez pomniejszenia zakresu rozwiązywanych przezeń zadań) zbliża nas do ważnego etapu naszej dyskusji: stworzenia najprostszego modelu komputera, który byłby formalnie równoważny wszystkim innym komputerom. Taki cel przyświecał Turingowi, gdy definiował swoją celebrowaną maszynę. Zastanówmy się, co to znaczy i co chcemy w ten sposób osiągnąć. Prostota modelu pozwoli nam go analizować przy pomocy narzędzi matematyki i produkować matematyczne, a więc niepodważalne, tezy. Ich głoszenie ma sens jedynie w kontekście abstrakcyjnych modeli, gdyż rzeczywistość jest zdradliwa i matematycy za nią (słusznie) nie przepadają. Gdybym chciał udowodnić matematycznie, że mój laptop potrafi policzyć jakąś skomplikowaną funkcję, stanąłbym przed problemem analizy astronomicznej wręcz liczby detali związanych z jego niezwykle pogmatwaną konstrukcją, której pełne ogarnięcie przez pojedynczego człowieka nie jest w dzisiejszych czasach w ogóle możliwe. Poza tym, rzeczywistość nas naprawdę zdradza. Jestem absolutnie pewien, że mój laptop zawiera błędy, zarówno sprzętowe (w elektronice) jak programowe (w softwarze). Prawdziwie dogłębna analiza jego możliwości (zakładając, że chcielibyśmy ją przeprowadzić) rozczarowałaby nas i doprowadziła donikąd. Zapamiętajmy: matematyk analizuje wyłącznie modele, gdyż rzeczywistość się do tego nie nadaje.

Równoważność modelu z rzeczywistymi komputerami posiada charakter abstrakcyjny, lecz reprezentatywny. Co to znaczy? Przede wszystkim nie interesuje nas różnica w czasie obliczenia. Jeśli jeden komputer potrafi rozwiązać dany problem w 10 sekund, podczas gdy drugiemu zajmie on 10 lat, ciągle powiemy, że (przynajmniej w przypadku tego konkretnego problemu), oba komputery są matematycznie równoważne. Przyspieszanie i usprawnianie czegoś, o czym wiadomo, że działa przebiega bowiem naturalną koleją rozwoju technologii (czego komputery konsekwentnie doświadczały przez ostatnie 80

lat) i nie ma w tym nic zaskakującego ani emergentnego. Gdy grupa matematyków z Bletchley Park,[63] w której osiągnięciach Alan Turing odegrał niemal tak prominentną rolę, jak odrobinę wcześniej Marian Rejewski i jego koledzy z polskiego Biura Szyfrów,[64] zaprezentowała swoje pokrętne (dosłownie oraz w przenośni) urządzenie do łamania kodu Enigmy Amerykanom, oni natychmiast usprawnili je przy pomocy sprytnych technicznych kruczków przyspieszając obroty bębnów i nieporównanie szybciej wykrywając ich interesujące konfiguracje. Entuzjazm matematyka kończy się na wykazaniu, że cel da się osiągnąć. Aby pokazać, jak go osiągnąć skutecznie, potrzebny jest inżynier.

Zmiana skali czasu jest oczywiście istotna dla ludzkiego użytkownika komputera, który pragnie jak najszybciej otrzymać wynik, lecz nie może mieć wpływu na jakościowe (wewnętrzne) aspekty obliczenia. Jeśli intryguje nas, czy komputer naprawdę myśli, czuje i posiada świadomość, to załóżmy przez chwilę, że tak jest. Jeśli jego subiektywny czas płynie szybciej lub wolniej niż nasz, ma to znaczenie jedynie z punktu widzenia naszych wzajemnych obserwacji świadomych poczynań drugiej strony nie dotykając ich subiektywnej natury. Jeśli przyśpieszymy powolny komputer tak, by zrównać nasze percepcje czasu, w świadomości komputera nic się nie zmieni. Zauważy on jedynie, że pewne zewnętrzne zjawiska, które dotąd zachodziły w przeraźliwie szybkim tempie stały się nagle ślamazarnie wolne. Podobnie stałoby się z nami, gdyby ktoś przyśpieszył nam zegary naszych myśli pozostawiając resztę świata bez zmian. Zmiana szybkości komputera, poza zmianą skali czasu, nie może zmienić natury jego obliczeń.



program

	0			1		
stan	wartość	akcja	stan	wartość	akcja	stan
1	1	L	2	1	P	1
2	0	P	3	1	N	7
3	1	P	5	0	P	9
...	...	...	...	...	...	...
9	0	HALT	---	1	R	8
10	0	L	6	0	L	5
...	...	...	...	...	...	...

Rysunek 16. Maszyna Turinga.

Wolno nam także założyć, że nasz abstrakcyjny komputer wyposażony jest w nieskończoną pamięć. Na pierwszy rzut oka można mieć obiekcje do równoważności takiego modelu z rzeczywistymi komputerami, ale założenie nieskończonej pamięci stanowi jedynie wygodny i akceptowalny skrót. W każdym przypadku, kiedykolwiek obliczenie komputera się zakończy (problem zostanie rozwiązany), ilość pamięci wykorzystana w obliczeniu będzie siłą rzeczy skończona (wykorzystanie nieskończonej ilości pamięci wymagałoby nieskończonego czasu). Tak więc celem założenia jest unikanie „biegania do sklepu" po dodatkową pamięć kiedykolwiek okaże się, że jest jej za mało dla rozwiązania aktualnego zadania. Model po prostu automatycznie dokupuje dodatkową pamięć, gdy okazuje się potrzebna.

[63] Sir F. H. Hinsley and Alan Stripp, Codebreakes: The Inside Story of Bletchley Park, Oxford University Press, 2001.
[64] Marian Rejewski, Wspomnienia z mej pracy w Biurze Szyfrów Oddziału II Sztabu Głównego w latach 1930-1945, LAM Wydawnictwo Naukowe, 2012.

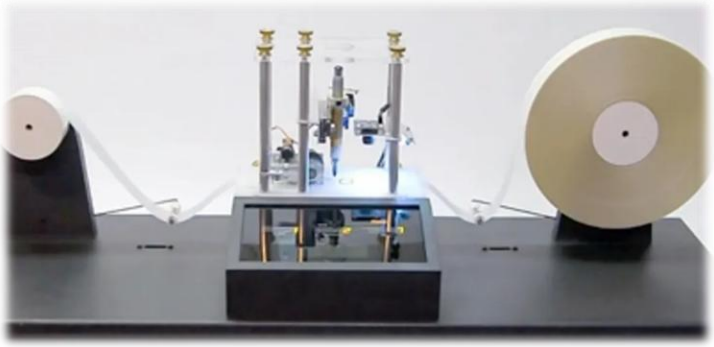
Maszyna Turinga (Rysunek 16) składa się z abstrakcyjnego procesora oraz abstrakcyjnej pamięci. W modelach wszystko jest abstrakcyjne, będziemy zatem pomijać ten przymiotnik w dalszej części naszego opisu. Nie mamy obowiązku wdawać się w szczegóły, lecz mimo to opiszmy maszynę Turinga dokładnie – by w pełni docenić jej prostotę. Jest to bowiem – pamiętajmy – najprostsza maszyna, która potrafi nie mniej niż najpotężniejszy komputer na świecie. Jeśli zatem zrozumiemy jej działanie, przestaniemy czuć się przytłoczeni ogromem komplikacji rzeczywistego sprzętu niezależnie od skali. Choćbyśmy sprzęgli ze sobą wszystkie komputery świata i kazali im rozwiązywać nie wiadomo jak zawiłe problemy, nie osiągniemy nic ponad to, co potrafi wyprodukować dziecinnie prosta maszynka o komplikacji liczydła. Jedyną różnicą jest czas oczekiwania na wynik. Ustaliliśmy, że nie może on mieć wpływu na istotę zjawisk zachodzących w trakcie obliczenia.

Pamięć maszyny Turinga posiada formę taśmy składającej się z przegródek, z których każda może przechować (zapamiętać) jeden z dwóch symboli, na przykład zero lub jedynkę. Maszyna wykonuje program opisany listą ponumerowanych napisów zwanych instrukcjami lub stanami. Procesor wyposażony jest w dwie strzałki: jedna, zwana głowicą, pokazuje na bieżącą przegródkę na taśmie, druga podaje bieżący numer stanu, czyli instrukcji do wykonania w danym kroku. Na początku, przed rozpoczęciem wykonywania programu, taśma zawiera zapis (reprezentację) problemu do rozwiązania, głowica wskazuje przegródkę taśmy z początkiem zapisu problemu, a numer stanu ustawiony jest na pierwszy stan z listy, czyli stan o numerze 1.

Wykonanie programu przebiega krokami w takt interpretowania treści stanów. Na początku kroku maszyna odczytuje z taśmy zawartość przegródki wskazywanej przez głowicę. Operacja opisana bieżącym stanem zawiera dwie części: jedną na okoliczność zera, drugą na okoliczność jedynki. Każda z części opisuje trzy akcje: co wpisać w bieżącą przegródkę taśmy (zero czy jedynkę), jak przesunąć głowicę i który stan uczynić stanem następnym (jak ustawić wskaźnik stanu przed następnym krokiem). Możliwości przesunięcia głowicy są trzy: jedna pozycja w lewo, jedna pozycja w prawo i nie przesuwać (pozostawić tam, gdzie jest). Numer następnego stanu musi się zawierać w zakresie liczby stanów programu. Jeden specjalny stan, o wybranym numerze i historycznej nazwie HALT, służy do zastopowania maszyny, czyli zakończenia programu. Możemy się umówić, że jest to ostatni stan na liście. Nie musi on zawierać żadnych operacji, gdyż po jego osiągnięciu maszyna się zatrzyma.

Dla zabawy możemy sobie wyobrazić fizyczną realizację maszyny Turinga w postaci bardzo długiego drutu z nanizanymi nań paciorkami. Każdy paciorek można obracać (przełączać) ruchem palca między dwiema pozycjami odpowiadającymi zeru i jedynce. Wyobraźmy sobie, że wykonujemy program zapisany na kartce przesuwając palec wzdłuż paciorków, wyczuwając stan paciorka znajdującego się aktualnie pod palcem, po czym zaglądamy w program i wykonujemy operację opisaną przez bieżący stan dla danej pozycji paciorka. Wszystko, co musimy pamiętać to numer bieżącego stanu uaktualniany po wykonaniu każdej instrukcji; możemy go sobie notować na boku. Kończymy zabawę, gdy osiągniemy stan HALT. Konfiguracja paciorków na drucie przedstawia wtedy wynik obliczeń.

Opisany powyżej aparat wygląda jak niestandardowe liczydło. Zamiast obsługiwać je palcem, dałoby się sprokurować niezbyt skomplikowany werk, gdzie zapadka wyczuwa pozycję bieżącego paciorka przesuwając mechanizm interpretujący zawartość stanu, który mógłby być opisany perforacją jakiegoś nośnika, na przykład papieru lub plastiku. Opis pojedynczego stanu mógłby stanowić rząd dziurek na nośniku zorganizowany w dwie grupy, dla zera i dla jedynki, i tak dalej. Jako niespełniony majsterkowicz o dwóch lewych rękach dostrzegam drobny problem z obsługiwaniem i dekodowaniem numeru stanu, ale znam ludzi, którzy wyśmieją moje objawy – nie takich rzeczy dokonywali w garażu w jedno popołudnie.



Rysunek 17. Jedna z fizyczna inkarnacji maszyny Turinga.[65]

W samej rzeczy, istnieje wiele fizycznych inkarnacji maszyny Turinga sprokurowanych przez zawziętych hobbystów (Rysunek 17). Formalnie, przy pomocy takiego urządzenia, potrafimy rozwiązać wszystkie problemy, jakie dają się opisać cyfrowo (jako sekwencje zer i jedynek) i dla których znamy procedurę rozwiązania zwaną algorytmem. Innymi słowy, maszyna Turinga potrafi policzyć wszystko, co potrafi policzyć najbardziej skomplikowany komputer. Dla zwykłego komputera, algorytm zapisany jest zwykle w jakimś uczciwym języku programowania, podczas gdy maszyna Turinga otrzymuje go w postaci listy opisów stanów. Nie jest to jednak poważny problem, gdyż dla każdego programu wyrażonego w dowolnym języku programowania, istnieje równoważny mu program maszyny Turinga, który można wyprodukować w procesie całkowicie mechanicznej translacji.

Mówiąc, że komputer rozwiązuje jakieś zadanie posługujemy się skrótem; powinniśmy raczej powiedzieć, że zadanie rozwiązuje program wykonywany na komputerze. Wszyscy rozumiemy, że uniwersalność współczesnych komputerów bierze się z łatwej wymienialności programów (software) przy niewymienialnym (a ściślej mówiąc znacznie trudniej wymienialnym) sprzęcie (hardware). Opisując maszynę Turinga możemy oczywiście oddzielić jej (umowny) sprzęt od (równie umownego) programu. Matematycy korzystają jednak nagminnie z tego samego potocznego skrótu nazywając maszyną Turinga pełną konfigurację, czyli sprzęt wraz z jej programem. Upraszcza to dyskusję, gdyż można wtedy powiedzieć, że dana maszyna rozwiązuje określony problem, albo że istnieje maszyna Turinga, która potrafi to lub tamto. Ma to sporo sensu, jako że surowy sprzęt maszyny Turinga jest banalny i niekontrowersyjny wśród specjalistów. Poza tym, matematyk może chcieć zmodyfikować go nieco, tak by interesujący program pozwalał się łatwiej wyrazić. Modyfikacje takie, podobnie jak różnice w komputerach dostępnych na rynku, nie zmieniają fundamentalnych własności wirtualnego urządzenia, lecz mogą uprościć matematyczne formuły czyniąc niektóre wywody bardziej klarownymi. W przypadku maszyny Turinga przeróbki sprzętu są łatwe, gdyż istnieje on jedynie w wyobraźni badacza; nie trzeba zatem biegać do sklepu i pozbywać się pieniędzy.

Pojawiają się zatem rozmaite wersje maszyny Turinga, na przykład posługujące się bogatszym zestawem symboli (alfabetem) niż zero i jedynka. Komplikując maszynę Turinga nie musimy się przejmować, że stworzymy coś, co wykroczy poza zakres możliwości oryginału, gdyż – pamiętamy – wszystkie komputery świata razem wzięte nie są w stanie wykroczyć poza ten zakres. Nawiasem mówiąc, nawet nasza prościutka wersja maszyny Turinga nie jest najprostszą z możliwych. Opcję „nie” dla operacji przesuwania głowicy można wyeliminować przyjmując, że jedynymi dopuszczalnymi opcjami są w lewo i w prawo. Jeśli w danym stanie maszyny głowica ma pozostać na bieżącym miejscu, wówczas możemy ją przesunąć w lewo, po czym przejść do nowego stanu, w którym głowicę przesuniemy w prawo dla obu możliwych zawartości przegródki wpisując tam poprzednią wartość, więc nie zmieniając nic. Utrudni

[65] https://spectrum.ieee.org/032610-diy-turing-machine . Konstruktorem urządzenia jest Mike Davey. Rolę taśmy wypełnia rolka białej folii, po której wymazywalny flamaster pisze zera i jedynki.

to trochę programowanie i powiększy liczbę stanów w programie, lecz – jeśli naszym celem jest uproszczenie abstrakcyjnego sprzętu maszyny – wolno nam to uczynić bez wpływu na jej formalne możliwości.

Próbując zastosować maszynę Turinga do programowania inteligentnych i świadomych umysłów (na przykład naszych) należałoby jeszcze zatroszczyć się o parę drobnych szczegółów. Co zrobić, na przykład, z naszymi zmysłami? Normalny komputer wyposażony jest w urządzenia peryferyjne, którymi mogą być sensory (pobierające informację z zewnątrz) i aktywatory (wyprowadzające informację na zewnątrz w postaci akcji wpływających na otoczenie). W przypadku maszyny Turinga, możemy umówić się, że pewne ustalone fragmenty taśmy dostarczają cyfrowego interfejsu do takich urządzeń, podobnie jak specjalne rejestry w przestrzeni pamięci normalnych komputerów. Zawartości rejestrów wejściowych mogą być częściowo stochastyczne odpowiadając niepewności informacji napływającej do nas ze świata. Ta stochastyczność może wpływać na akcję programu odzwierciedlając niepewność i indeterminizm naszych decyzji.

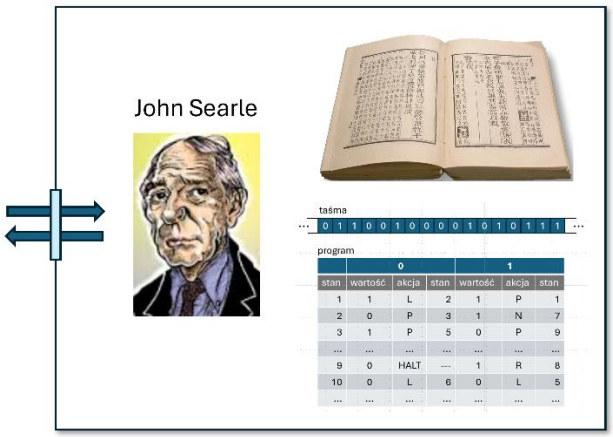
Sprowadzalność wszelkich akcji kiedykolwiek i gdziekolwiek dokonywanych przez jakiekolwiek komputery do kroków maszyny Turinga posiada nietuzinkową siłę demistyfikacji ich możliwości wykorzystywaną przez niektórych filozofów. Mówiąc brutalnie: liczydło nie myśli. Maszyna Turinga, będąc równoważna wszystkim komputerom, obecnym i przyszłym, jest także równoważna liczydłu o nieco nietypowej konstrukcji, którego ideę naszkicowaliśmy powyżej. Skąd w takim banalnym urządzeniu miałaby się wziąć świadomość, cokolwiek by to nie było? Jasne, że liczydło nie myśli leżąc sobie bezczynnie. No ale dołożenie mu dynamiki w postaci mechanicznego palca obracającego paciorki na drucie nie zakrawa na postęp prowadzący do niespodziewanej „emergencji" świadomości. By oddać honor maszynie Turinga, palec musi być wyposażony w werk wyczuwający położenie paciorka oraz perforacje na nośniku z programem i przesuwający palec wzdłuż drutu. Zauważmy jednak, że zawartość programu (perforacji) nie zmienia się w trakcie jego wykonywania, więc na tym odcinku absolutnie nic się nie dzieje. To prawda, że dla poważnego programu rozmiar nośnika z perforacjami może okazać się astronomiczny, lecz co z tego? Czy góra kamieni posiada większą inteligencję niż pojedynczy kamień, szczególnie jeśli leży sobie spokojnie bez ruchu? Dla dowolnego rzeczywistego programu, na przykład takiego, który realizuje sieć neuronową olbrzymiego modelu języka, potrafimy podać dokładny przepis wyprodukowania perforacji równoważnego programu dla maszyny Turinga. Potem uruchomimy nasz mechaniczny palec i model ożyje. Czy uzyska świadomość?

Problem polega na tym, że zupełnie nie widać miejsca, w którym świadomość mogłaby się ujawnić, czyli dokonać swojej emergencji. Cała bowiem dynamiczna komplikacja systemu skupiona jest w banalnym mechanizmie, który wygląda tak samo dla programu kręcącego się w miejscu przestawiając w nieskończoność pojedynczy paciorek, jak dla modelu języka odbywającego właśnie test Turinga. Jeżeli gdzieś w tym prymitywnym zestawie drążków, trybików i być może kabelków znajduje się siedlisko ducha, to należy go poszukiwać w programie i tworzonym przez niego zapisie, ale – jak zauważyliśmy – program się nie zmienia, a zapis modyfikowany jest drobnymi kroczkami, których zasięg dotyczy opuszka palca dotykającego bieżącego paciorka.

Uderzająca jest także przyrodzona prostota konstrukcji sieci neuronowej, nawet jeśli rozważymy ją w oderwaniu od komputera (równoważnego maszynie Turinga), na którym jest ona zwykle realizowana. Jej inspiracją i formalnym odpowiednikiem jest gromada potencjometrów połączonych drucikami, czyli dość prostacka (aczkolwiek masywna w przypadku modelu języka) konfiguracja urządzeń całkowicie banalnych. Określenia w rodzaju „neuron", „długa pamięć krótkoterminowa", „skupienie uwagi" i szereg innych, jeszcze bardziej sugestywnych terminów używanych i nadużywanych przez profesjonalistów mogą brzmieć tajemniczo, sugerując mistyczne moce, co nie zmienia faktu, że każdy z tych zaawansowanych komponentów wypełnia żałośnie beznamiętną funkcję.

Cóż więc nadaje mechanizmowi zdolność introspekcji, refleksji, zadumy, niepokoju, podniecenia? Na którym etapie usprawniania liczydła przez podstawianie mu coraz większej góry kamieni przekraczamy granicę oddzielającą bezduszny złom od natchnionego umysłu?

Filozof John Searle ilustruje problem eksperymentem myślowym przybranym w barwną dykteryjkę znaną pod nazwą Argumentu Chińskiej Komnaty.[66] W zamkniętym pokoju, do którego jedyny dostęp stanowi wąskie okienko, przebywa człowiek, na przykład John Searle (Rysunek 18), wyposażony w tekst programu (listę stanów maszyny Turinga), który jest w stanie zaliczyć test na inteligencję konwersując w języku chińskim. Dodatkowo w pomieszczeniu znajdują się tablice kodowania i dekodowania chińskich znaków na sekwencje zer i jedynek. John Searle odgrywa rolę mechanizmu wykonującego program. Pobiera przez okienko zapytanie w postaci ciągu chińskich symboli, zamienia go przy pomocy tablic na dane dla maszyny Turinga, następnie krok po kroku wykonuje wszelkie wymagane operacje manualnie przestawiając paciorki zgodnie z instrukcjami programu. Na koniec dekoduje zestaw zer i jedynek wyprodukowany przez maszynę na ciąg chińskich symboli zwracając go przez okienko jako wynik. Program udziela inteligentnych odpowiedzi na zadane mu pytania, zatem zachodzi podejrzenie, że myśli i rozumie treść przeprowadzanych z nim konwersacji. John Searle twierdzi, że jest to absolutnie niemożliwe, gdyż wszystko, co się dzieje w pomieszczeniu, to przesuwanie kartek i paciorków jego własną ręką, których to czynności nie sposób podejrzewać o myślenie czy rozumienie czegokolwiek. Na dodatek on, John Searle, nie zna ani słowa po chińsku, zatem te wszystkie symbole pobierane i przekazywane przez okienko stanowią dla niego ... chińszczyznę, której on rzecz jasna nie jest w stanie pojąć.



Rysunek 18. Chińska Komnata.

Argument Chińskiej Komnaty nie wymaga nawet maszyny Turinga. W końcu każdy komputer, choćby najbardziej skomplikowany, realizuje pewien ustalony repertuar operacji (instrukcji maszynowych) na sekwencjach bitów. Rzeczone operacje posiadają oczywiście precyzyjną dokumentację, z której korzystają programiści. Człowiek jest zatem w stanie emulować ich wykonanie przy pomocy ołówka i kartki, notując stany pamięci przed i po ich wykonaniu. Teoretycznie można w ten sposób realizować dowolnie złożony program (algorytm) dla dowolnych danych bez fizycznego komputera. Koncepcja algorytmu jako formalnej procedury wykonywanej na liczbach lub symbolach znana jest co najmniej od czasów Euklidesa, kiedy o komputerach nikomu się nie śniło. Sam Turing projektował algorytmy dla zespołu rachmistrzów zatrudnionych w Bletchley Park (zwanych z anglosaska komputerami), którzy mechanicznie wykonywali żmudne obliczenia, często bez znajomości celu kompletnego programu. W początkowym okresie moich studiów informatycznych, gdy o komputerach osobistych nikt nie słyszał, a

[66] John Searle, Minds, Brains, and Programs, The Behavioral and Brain Sciences 3, pp. 417-424, 1980.

wydziałowa maszyna cyfrowa zajęta była poważniejszymi sprawami niż wspieranie raczkujących studentów, ręczne wykonywanie edukacyjnych programików znajdowało się na porządku dziennym.

Podsumowując, realizacja programu komputerowego polega na „bezmyślnym" wykonywaniu prostych mechanicznych kroków, które model Turinga, formalnie równoważny wszystkim komputerom, sprowadza do przesuwania paciorków liczydła. Cała złożoność systemu obliczeniowego zawarta jest w statycznym programie, który może być wiernie reprezentowany tekstem zapisanym na kartce papieru. W samej rzeczy, programy pisano niegdyś na papierze, potem przepisywano je na tasiemki lub karty perforowane, które następnie wczytywano w komputer. Przekształcenie programu zakodowanego na kartach na użytek jego interpretacji przez komputer stanowiło prosty i automatyczny zabieg zmieniający jedynie formę przy wiernym zachowaniu treści. Można sobie wyobrazić obliczenie, w którym karty perforowane (zawierające zresztą, poza dziurkami, czytelny tekstowy zapis kodowanych przez nie instrukcji) są interpretowane przez człowieka wykonującego zapisany na nich program bez zrozumienia danych oraz wyniku. Pytanie skąd w takim systemie ma się pojawić świadomość jest, przynajmniej dla niektórych filozofów, intrygujące.

Użycie człowieka jako procesora wykonującego program w eksperymentach myślowych Searla (i jemu podobnych) jest, przynajmniej na mój gust, niezbyt potrzebne. Ludzki procesor uwypukla problem podkreślając absurdalność tezy, że coś w tym hybrydowym komputerze rozumie chiński. Skoro bowiem jedyny niepodważalnie myślący element systemu nie ma pojęcia co robi, a na nim skupia się cała dynamika obliczenia, to skąd ma się tam wziąć myślenie, zrozumienie i świadomość? No ale zastąpienie człowieka mechanicznym procesorem zbudowanym z drążków, trybów, zapadek, elektromagnesów i drutów (czy formalnie równoważnym zestawem układów scalonych współczesnego komputera) nie zmienia przecież istoty problemu.

Argument Chińskiej Komnaty, że myślenie i świadomość nie pozwalają się sprowadzić do algorytmu, nie jest w swojej esencji specjalnie świeży. Gottfried Leibniz, w swojej słynnej Monadologii,[67] w roku 1714 pisze (moje tłumaczenie):

> „Należy uznać, że postrzeganie i wszystko co z niego wynika nie da się wytłumaczyć na gruncie mechaniki, czyli przez przedmioty oraz ich ruch. Założywszy, że istniałaby maszyna skonstruowana w ten sposób, by myśleć, postrzegać i czuć, można sobie wyobrazić jej konstrukcję o powiększonych rozmiarach, przy zachowaniu proporcji, do której potrafilibyśmy wkroczyć niczym do młyna. Zbadawszy jej wnętrze, znaleźlibyśmy tam jedynie części poruszające inne części i nic, co wytłumaczyłoby percepcję. Zatem, percepcji należy poszukiwać w prostej substancji, a nie w substancji złożonej lub w maszynie."

Nie mamy tu miejsca na poważną dygresję w stronę Monadologii Leibniza. Poprzestańmy zatem na objaśnieniu, że przez percepcję Leibniz rozumie świadomość, zaś prosta substancja, w której należy jej poszukiwać, jest czymś zgoła odmiennym od tego, co współczesny naturalistyczny świat rozumie przez materię. Tak więc konkluzja Leibniza brzmi, że (zwykła) materia nie jest w stanie dostarczyć bazy dla zaistnienia świadomości. Nie może zatem wyprodukować jej algorytm realizowany na komputerze.

Argumenty nie stanowią matematycznego dowodu. Nie zbliżają nas do niczego, co mogłoby stać się choćby zalążkiem fizycznej teorii świadomości, gdzie argumentacja posługiwałyby się matematyką. Nie wiemy, czym dokładnie jest świadomość, więc nie potrafimy wyartykułować warunków koniecznych dla jej pojawienia się. Teza Churcha-Turinga[48] spekuluje, że wszystkie procesy fizyczne, cały dynamizm naszego świata, dają się obliczać algorytmicznie, czyli że nie istnieje nic, czego maszyna Turinga nie potrafiłaby emulować. Mówiąc dokładniej, głosi ona, że maszyna Turinga (a zatem komputer) potrafi

[67] Nicholas Rescher, G. W. Leibniz's Monadology: An Edition for Students, University of Pittsburgh Press, 1991.

policzyć wszystko, co da się w ogóle policzyć efektywnie, czyli w fizycznej rzeczywistości. Jeśli nasza świadomość tam przebywa, wynikałoby z tego, że znajduje się ona w zasięgu komputera – czyli liczydła.

Stanowisko, jakie zajmuje Searle nie jest zgodne z tezą Churcha-Turinga. Uważa on, że algorytm realizowany na komputerze nigdy nie posiądzie autentycznego zrozumienia i świadomości, gdyż wszystko, co potrafi komputer to przetwarzanie składni (sekwencji symboli). Dla rozumienia, myślenia i świadomości potrzebna jest semantyka, która nigdy nie wyniknie z samej składni przez stosowanie do niej mechanicznych reguł. Nie uciekając się do mistycyzmu ani religii i nie próbując objaśniać fenomenu świadomości, Searle konkluduje, że systemy biologiczne funkcjonują według zasad, które nie są obliczalne w sensie Churcha-Turinga. Procesy biologiczne nie są zatem w pełni objaśnione mechanistyczną wersją współczesnej fizyki i chemii, gdyż znajdują się tam zjawiska falsyfikujące rzeczoną tezę. Ma z tego zatem wynikać, że rzeczywistość nie jest obliczalna. Uspokajającym pobocznym efektem tej konkluzji jest zdementowanie pogłosek jakobyśmy odgrywali role awatarów w grze komputerowej uprawianej przez dzieci z innego kosmosu.

Pomimo starannego unikania mistycyzmu, Searle'owi udało się poirytować zwolenników tak zwanej twardej sztucznej inteligencji oraz niektórych filozofów o mniej spolaryzowanych poglądach, którzy próbują dostrzegać luki w jego rozumowaniu.[68] Ci pierwsi twierdzą, że w systemie biologicznym (na przykład człowieka) nie dzieje się nic takiego, czego komputer nie potrafiłby odtworzyć przy pomocy odpowiednio zmyślnego programu. Drudzy, niekoniecznie zgadzając się z pierwszymi, uważają, że argumenty Searle'a nie są dostatecznie przekonujące. Patrick Hayes, prominentny wyznawca twardej sztucznej inteligencji zakładającej policzalność absolutnie wszystkich aspektów ludzkiego rozumu, zredefiniował cel badań dyscypliny jako refutację argumentu Searle'a.[69] Reakcje takie wykazują spory ładunek zakładanej z góry ideologii ustawiającej powóz w postaci wyniku badań przed koniem, czyli badaniami, które mają do niego (rzecz jasna obiektywnie) prowadzić. Trudno tego uniknąć w tych spośród naukowych dysput, gdzie autorytatywne demonstracje nie są z zasady możliwe i gdzie granice ideologii dotykają granic światopoglądowych.[57] Zauważmy, że aby wysunąć się poza zakres demagogii określonej emocjonalnym traktowaniem tautologii Turinga, refutacja Argumentu Chińskiej Komnaty musiałaby zawierać precyzyjną definicję świadomości, na którą zgodziliby się matematycy i filozofowie. Oczywiście ChatGPT nie dostarcza takiej refutacji, co warto podkreślić, biorąc pod uwagę, że postulat Hayesa ma już ponad 30 lat.

Ataki na Argument Chińskiej Komnaty przebiegają wzdłuż kilku linii, które dla niewyrobionego filozoficznie hobbysty (jak ja) są mało odróżnialne. Według jednej z nich, praktyczne udawanie przez człowieka procesu obliczeniowego współczesnego superkomputera nie jest możliwe ze względu na różnicę skali czasowej. Udzielenie odpowiedzi na jedno pytanie trwałoby bowiem tysiąclecia.[70] Przy emulowaniu myślącego komputera okazałoby się, że system składający się z Johna Searle i wszystkich wymaganych rekwizytów jego Chińskiej Komnaty faktycznie myśli i posiada świadomość, co przejawia się w sposób niezauważalny w skali postrzeganej przez człowieka, czyli w tempie przekładania kartek i kreślenia notatek na papierze. Dlatego nie powinniśmy się dziwić, że nie dostrzegamy gołym okiem miejsca, gdzie świadomości należy szukać. Ona się tam dzieje, lecz bardzo, bardzo powoli.

Istnieje klasa obiekcji przeciwko modelom i eksperymentom myślowym, które będąc formalnie zgodne z prawami fizyki nie są realizowalne w rzeczywistości, na przykład ze względu na brak dostatecznej ilości zasobów w namacalnym świecie lub, jak w tym przypadku, niemożność zaobserwowania zjawiska w skali czasowej pojedynczego człowieka. Orędownicy takich obiekcji zakładają istnienie czegoś w

[68] The Chinese Room Argument, Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/ENTRIES/chinese-room/ .
[69] Patrick Hayes et. al, Virtual Symposium on Virtual Mind. Minds and Machines 2, pp. 217-238, 1992.
[70] Frank Tipler, The Physics of Immortality, Anchor, 1997.

rodzaju cenzury kosmicznej, która torpeduje filozoficzne wnioski wynikające z formalnych praw fizyki oraz reguł matematyki, jeśli posuwają się one zbyt daleko przeciwko wytycznym cenzora. Wpływ skali czasowej na stosowalność Argumentu Chińskiej Komnaty byłby jedną z konsekwencji takich ograniczeń. Na mój gust, oznaczałoby to konspirację nieznanych sił (trudno je nazwać mechanizmami), przy których sam Argument Chińskiej Komnaty wydaje się domagać delikatniejszych ustępstw. Poza tym, jak zauważa Searle, jego ludzki komputer można przyspieszać, na przykład zatrudniając wszystkich mieszkańców Chin jako procesory (tzw. Argument Chińskiego Narodu).[68] Równoważność modelu obliczeń Turinga dopuszcza oczywiście wieloprocesorowe przetwarzanie równoległe bez wprowadzenia nowej jakości w proces obliczenia.

Głównym hasłem krytyków Searle'a jest jednak system. Oczywiście zgadzają się oni z Searle'm, że nie rozumie on nic z wykonywanych przez siebie funkcji, lecz podkreślają, że ludzki procesor jest jedynie fragmentem większego systemu, który – jako całość – ma prawo myśleć i posiadać świadomość (o czym Searle jako jeden z jego komponentów nie musi wiedzieć).[71] Moim skromnym zdaniem, Searle wystawił się na taki typ kontrargumentów nalegając, by procesorem był człowiek, co – jak zauważyliśmy wcześniej – nie zmienia istoty problemu przystrajając ją jedynie płaszczykiem prowokującej narracji. Chwyt odegrał swoją (ewidentnie zamierzoną) rolę popularyzacji argumentu, który znany był filozofom co najmniej od czasów Leibniza. Znajoma nam piosenka z minionej epoki uczy, że nic tak nie pobudza uczonych do działania, jak telewizja sugerująca im czym w danej chwili powinni się zająć.

Teza krytyków Argumentu Chińskiej Komnaty jest zatem taka, że komputer wykonujący program potrafi myśleć i doznawać świadomości na zasadzie bycia złożonym systemem zbudowanym z paciorków liczydła przesuwanych ludzkim bądź mechanicznym palcem w oparciu o statyczną listę instrukcji zapisanych na kartce, plastikowej tabliczce lub innym nośniku. Krótko mówiąc – nie ma problemu, prosimy się rozejść, świadomość w systemie jest, gdyż być musi. Systemy tak już mają, że dzieją się w nich zjawiska złożone. Całość to czasem więcej niż suma części. No i przecież – wiadomo – emergencja.

Nie do końca dla mnie jasne (i jak rozumiem drobne) nieporozumienie wśród wyznawców emergencji świadomości w odpowiednio dużym liczydle dotyka jednak kwestii kto (czy co) dokładnie tam myśli. Jedni twierdzą, że po prostu system jako taki, drudzy, że system wytwarza coś na kształt wirtualnego homunkulusa i że to on dopiero uprawia rzeczoną działalność.[72] Jako programista, który w życiu napisał i zobaczył więcej prawdziwych (i dużych) programów niż typowy teoretyk informatyki (a z pewnością znacznie więcej niż typowy filozof) nie potrafię przejść do porządku dziennego nad teorią umysłu rodzącego się samoczynnie między palcem a obracanymi przezeń paciorkami. Nie potrafię, tym bardziej że na własnych nogach przebyłem drogę od entuzjastycznej wiary w algorytmiczną potęgę programów do osobistej konstatacji tego, co geniusz Leibniza dostrzegł 250 lat przed współczesnym sformułowaniem problemu: jak by na to nie patrzeć, widać tam wyłącznie paciorki i poruszający je palec.

Nie wszyscy fizycy i matematycy, uprawiający „twardą" naukę, przepadają za wsłuchiwaniem się w opinie filozofów. Richard Feynman, noblista i współtwórca elektrodynamiki kwantowej, stwierdził kiedyś, że filozofowie są tak użyteczni naukowcom, jak ornitolodzy ptakom. Roger Penrose, także noblista, jeden z nielicznych fizyków/matematyków zajmujących się na poważnie problemem świadomości, docenia wagę argumentu Leibniza poszukując dróg rozwiązania zagadki w nowych i nieortodoksyjnych teoriach fizyki. Jego argument przeciwko obliczalności myślenia[73] i świadomości jest bardziej formalny niż dykteryjka Searle'a, lecz nie mniej kontrowersyjny. Penrose zauważa, że celebrowane twierdzenie

[71] Georges Rey, What's Really Going on in Searle's "Chinese Room", Philosophical Studies, 1;50(2), pp. 169-185, 1986.

[72] Jack B. Copeland, Logical Point of View, Views into the Chinese Room: New essays on Searle and artificial intelligence, 109, 2002.

[73] Roger Penrose, Shadows of the Mind, Oxford University Press, 1994.

Gödla o nierozstrzygalności formalnych systemów wnioskowania zawierających arytmetykę liczb naturalnych,[74] czyli na przykład programów komputerowych, da się obejść przez system zdolny do introspekcji, a więc wyposażony w świadomość (podobną obserwację poczynił wcześniej John Lucas).[75] Wynikałoby z tego, że myślący program, którego logika musi ściśle przestrzegać twierdzenia Gödla, nie może być zdolny do introspekcji, zatem nie może być świadomy. Argument Lucasa-Penrose'a słabnie z braku możliwości sformalizowania świadomości (introspekcji), co sprawia, że wszystko sprowadza się znów do subiektywności postrzegania. Argument, który Penrose (świadomie) wygłasza (werbalnie bądź na papierze) może bowiem wyprodukować komputer zaprogramowany specjalnie po to, żeby go właśnie wygłosić albo wydrukować. Mamy zatem słowo Penrose'a przeciwko słowu wyprodukowanemu przez komputerową drukarkę. Nie sposób udowodnić, że drukarka nie dokonała introspekcji drukując „swój" argument, no bo jak to uczynić, jeśli matematyka nie wie co to introspekcja. I tak w kółko.

Siła argumentu Penrose'a jest zatem subiektywna, podobnie jak w przypadku argumentów Leibniza i Searla oraz całej masy podobnych argumentów wygłaszanych przez myślicieli różnych maści zaniepokojonych tajemnicą świadomości i jej intuicyjną niezgodnością z fenomenem mechanicznego obliczenia. Moje własne przemyślenia w tym zakresie zaczęły się dość późno. Na początku edukacji, młody człowiek, jeśli przejawia zdrową dozę entuzjazmu do wiedzy, podlega fascynacji świeżo nabytą mądrością i łatwo mu uwierzyć w jej bezgraniczną potęgę. Czasem taka wiara pozostaje na dłużej, szczególnie jeśli specjalizujemy się w dziedzinie, w którą wierzymy i od naszego entuzjazmu zależy energia jaką wkładamy w badania i w konsekwencji nasz sukces w zdobywaniu popularności i środków wspierających nasz rozwój. Trudno uważać to za coś złego; wręcz przeciwnie – uczciwa fascynacja, nawet zwodnicza, stanowi główny motor postępu w nauce.

Mój prywatny model obliczeń, stanowiący niegdyś powracający temat nocnych koszmarów, skonstruowany był z pociągów towarowych przewożących kamienie, przemieszczających się po gmatwaninie torów, platform i zwrotnic. Śniłem te pociągi rozumiejąc, że realizują świadome obliczenie, gapiłem się w zwrotnice pośród upiornego turkotu metalowych kół i przyzywałem homunkulusa, który by mi wytłumaczył, co się w tym wszystkim dzieje. Można i tak. Modeli formalnie równoważnych maszynie Turinga (i w konsekwencji wszystkim komputerom) jest wiele i łatwo je wymyślać przy odrobinie wprawy. Pozorna przewaga takich, które służą czemuś innemu niż udawanie liczydła zasadza się na amplifikacji absurdu. Searle sprawił, że Chiny stały się ulubioną pożywką ich twórców. Oto przykład: cała populacja Chin emulująca odpalanie neuronów w ludzkim mózgu przez telefonowanie według ustalonego schematu.[76]

Dla głosicieli obliczalnego fizykalizmu, w którym obowiązuje teza Churcha-Turinga, świadomość jest cokolwiek niewygodna, gdyż „chińskie" argumenty przemawiają do każdego, choć w nierównym stopniu. Niektórzy więc, w obronie swoich pozycji, posuwają się do tezy, że świadomość zwyczajnie nie istnieje. Nie istnieje także wolna wola oraz parę innych rekwizytów. Dla niektórych, jak na przykład dla mnie, negowanie istnienia tych wszystkich nieidentyfikowalnych mechanistycznie elementów składających się na esencję naszego człowieczeństwa jest cokolwiek irytujące. Cytując Wolanda z „Mistrza i Małgorzaty",[77] czego byś nie tknął, tego najzwyczajniej nie ma.

Negację roli świadomości można uprawiać z grubsza na dwa sposoby. Wspomnieliśmy już o iluzjonizmie, który wywodzi, że świadomość zasadniczo nie istnieje stanowiąc złudzenie, coś w rodzaju snu, w

[74] Kurt Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme, I, Monatshefte für Mathematik und Physik, (38) 1, pp. 173–198, 1931.
[75] John Lucas, Mind, Machines, and Gödel, Philosophy, (36), pp. 112-127, 1961.
[76] Ned Block, Troubles with Functionalism, The Language and Thought Series, Harvard University Press, pp. 268-306, 1980.
[77] Michał Bułhakow, Mistrz i Małgorzata, MUZA, 2011.

którym dzieją się zwidy, podczas gdy prawdziwy materialny świat funkcjonuje całkiem normalnie nic sobie z tego nie robiąc. Nieco odmienny pogląd – zwany epifenomenalizmem – głosi, że wszelkie nasze ludzkie atrybuty, których nie da się dostrzec w mechanizmach to przypadkowe, poboczne zjawiska niegodne naukowych studiów.[78] Według niego, świadomość, introspekcja, poczucie wolnej woli, i cały ten bałagan, to jak syk pary w lokomotywie – nieważne, akcydentalne i bez związku z istotą rzeczy. Przypomina mi się scena z filmu „Manhattan" Woody'ego Allena,[79] gdzie podczas spaceru ulicami Nowego Yorku Diane Keaton oświadcza: „Miałam orgazm, ale mój psychoanalityk twierdzi, że nieprawdziwy".

Jednym z prominentnych wyznawców epifenomenalizmu jest Douglas Hofstadter, autor serii popularnych i celebrowanych (przynajmniej w strefie anglosaskiej) książek,[80] z których (o ile mi wiadomo) żadna nie doczekała się dotąd polskiego tłumaczenia. W jednej z nich[81] Hofstadter z właściwą mu poetycką erudycją opisuje swoją osobistą mieszankę pogardy i fascynacji fenomenem świadomości, który ma według niego stanowić poboczny efekt „miriad pętli" programu realizowanego przez neurony ludzkiego mózgu. Naturalistycznym przeznaczeniem rzeczonego mózgu jest coś zgoła innego, lecz ta nieszczęsna świadomość wypełza z niego szczelinami spowodowana astronomiczną liczbą pętli, w jakie zorganizowany jest jego program. Powoduje to rozpaczliwe problemy, z którymi człowiek zmuszony jest się zmagać w swoim skądinąd beznamiętnym istnieniu. Hofstadter napisał książkę po śmierci żony jako próbę dojścia do ładu z finalnością materialnego ustania najbliższej mu osoby. Jako ewangeliczny naturalista nawiguje w kierunku pokrętnego pseudofizykalnego mistycyzmu doszukując się ekwiwalentów (nieistniejącej z założenia) duszy w czeluściach tajemniczego, abstrakcyjnego komputera.

Wypadnie dla porządku zauważyć, że mierzenie komplikacji (czy stopnia świadomości) algorytmu liczbą zawartych w nim pętli ma tyleż sensu, co szacowanie inteligencji góry kamieni przez jej wysokość. Elementarną konsekwencją tej samej matematyki, która czyni wszystkie komputery równoważnymi maszynie Turinga jest bowiem fakt, że dowolnie skomplikowany program można zawsze (i całkowicie mechanicznie) przekształcić tak, że wystąpi w nim co najwyżej jedna pętla. Podobnie jak liczydło jest formalnie w stanie zrealizować to samo obliczenie co farma najbardziej zaawansowanych superkomputerów, tak miriada pętli najpokrętniej pogmatwanego programu da się bez większego trudu zastąpić jedną.

Kultowy i nieżyjący już amerykański popularyzator matematyki, Martin Gardner, w swojej recenzji książki Hoftstadtera pisze:[82]

> „Wyłożę karty na stół. Należę do niewielkiej grupy myślicieli zwanych „misterianami". Są wśród nas tacy filozofowie jak Searle (w książce Hofstadtera napiętnowany jako hochsztapler), Thomas Nagel, Colin McGinn, Jerry Fodor, a także Noam Chomsky, Roger Penrose i paru innych. Wszyscy podzielamy przekonanie, że żaden z żyjących dziś filozofów i naukowców nie posiada najbledszego wyobrażenia o tym, jak świadomość i jej nieodłączna towarzyszka wolna wola wyłaniają się (co ewidentnie ma miejsce) z materialnego mózgu."

I kończy swoją recenzję tak:

[78] Sven Walter, Epiphenomenalism, Internet Encyclopedia of Philosophy, https://iep.utm.edu/epipheno/ .
[79] Woody Allen, „Manhattan", 1979, https://www.imdb.com/title/tt0079522/ .
[80] Najpopularniejszą z nich jest: Douglas Hofstadter, Gödel, Escher, Bach: an Eternal Golden Braid, Basic Books, 1971.
[81] Douglas Hofstadter, I am a Strange Loop, Basic Books, 2007.
[82] Martin Gardner, Do Loops Explain Consciousness? Review of I am a Strange Loop, Notices of the American Mathematical Society, 54(7), pp. 852-854, 2007 (moje tłumaczenie).

> „Być może gdzieś w Mgławicy Andromedy zamieszkują zaawansowane formy życia znające odpowiedzi na nasze pytania. Ja ich z pewnością nie znam. Nie znają ich także Hofstadter i Dennet.[83] I nie znasz ich ty."

Czytelnik poszukujący ulubionej teorii świadomości, do której mógłby komfortowo zasubskrybować w zgodzie ze swoim światopoglądem, posiada rozległy wybór. Robert Lawrence Kuhn pokusił się niedawno o zebranie wszystkich mniej lub bardziej znanych idei na ten temat w jednym (olbrzymim) artykule, gdzie możemy się doliczyć około 250 niekompatybilnych spekulacji.[84] Nie są to niestety teorie naukowe, gdyż nie można ich formalnie przeciwstawiać sobie: wyznaczać predykcje, planować eksperymenty, dokonywać weryfikacji. Subskrypcja może się zatem dokonać jedynie na zasadzie subiektywnego odczucia, że któraś z rzeczonych spekulacji przypada nam do gustu lepiej niż inne. Pod koniec artykułu Kuhn cytuje Jerry'ego Fodora (o którym wspominaliśmy wcześniej):

> "Nikt nie ma bladego pojęcia, jak cokolwiek materialnego może być świadome. Nikt nawet nie wie, jak sobie wyobrazić, że się ma blade pojęcie, jak coś materialnego jest świadome."[85]

Jako człowiek związany z nauką nie ośmielam się nikomu narzucać mojego poglądu na kwestie, co do których nie posiadam dowodów. Ale rozumiem też, że twarde dowody można wytaczać jedynie w matematyce, podczas gdy w życiu … wszystko może okazać się iluzją. Wszystko zatem, przy czym mam prawo się upierać to moje luźne poglądy i opinie. Jedna z nich jest taka, że nie jestem komputerową emulacją. A już na pewno nie jestem emulacją komputera równoważnego maszynie Turinga. Możemy się umówić, że pośród zalewu metafizycznych spekulacji na temat „teorii" świadomości, jest to jedyne (binarne) odróżnienie, które nas interesuje w konfrontacji ze sztuczną inteligencją.

Załóżmy jednak, że się mylę i popatrzmy, dokąd nas to założenie doprowadzi. Przyjmijmy, że nie ma w nas nic osobliwego, że jesteśmy obliczalni i funkcjonalnie tożsami z programami maszyny Turinga. Wykonanie takiego programu to sekwencja ruchów głowicy i zapisów na taśmie rzetelnie opisująca absolutnie wszystko, co się w maszynie dzieje. Możemy ją zakodować na wąziutkim i długim pasku papieru i może ona wyglądać tak: …P1L0L1L0P1N0P0P1…, gdzie literki oznaczają ruch (prawo, lewo, nie), a po literce następuje zero lub jedynka oznaczająca nową zawartość przegródki wskazywanej przez głowicę. Taki zapis utrwali cały proces obliczenia, pełną informację wyprodukowaną przez maszynę i jej program, coś co programiści nazywają śladem.

Wyobrażam sobie, że trzymam w ręku hyperastronomicznej długości pasek papieru ze śladem świadomego programu. Efekt wykonania programu przez maszynę był w pełni równoważny pojawianiu się kolejnych elementów napisu pod bieżącą pozycją głowicy. Patrzę na samiutki koniec paska i widzę ostatni wpis, po którym maszyna wykonała HALT. Jeśli teraz w zatrzymanej i już martwej maszynie cofnę wskaźnik stanu o jedną pozycję, poprawię program i znów go puszczę w ruch, to czy ktoś w jakimś świecie zmartwychwstanie?

Cóż takiego właściwie trzymam w ręku? Jeśli zacznę przesuwać po pasku palec wskazując na sukcesywne znaki lub wypowiadać je po kolei, to czy przywołam do życia myślącą istotę, której świadomość zaszyta jest w tym martwym tekście? Czy moje słowo stanie się ciałem? To raczej oczywista bzdura. A co się stanie, jeśli zamiast przesuwania palcem po pasku przetworzę ślad przy pomocy prościutkiego

[83] Gardner wymienia tu nazwisko Denetta, który popełnił głośną swojego czasu książkę pod tytułem „Świadomość Wyjaśniona" (Daniel C. Dennet, Consciousness Explained, Bay Books, 1992). Tytuł książki to rodzaj bezczelnej przynęty marketingowej. Gdy po raz pierwszy wpadł mi w oczy, w księgarni „Borders" w Sacramento (Kalifornia), chwyciłem książkę i nerwowo przekartkowałem ją na miejscu, praktycznie do końca, wypijając przy tym wiaderko kawy i oczekując obiecanego „wyjaśnienia". Nietrudno zgadnąć, że go nie znalazłem.

[84] Robert Lawrence Kuhn, A landscape of consciousness: Toward a taxonomy of explanations and implications, Progress in Biophysics and Molecular Biology, 190 (2024), pp. 28-169, 2024.

[85] Moje tłumaczenie.

programu, który wykona zawarty w nim przepis wpisując w taśmę maszyny Turinga po kolei wszystkie wartości, które uprzednio produkował oryginalny program? Niezależnie jak monstrualny i niebotycznie skomplikowany był tamten świadomy program przeżywający swoje wzloty, namiętności, zwątpienia i upadki, mój obecny odtwarzacz śladu jest prostszy niż magnetofon kasetowy z ubiegłego wieku, lecz jego dynamiczny efekt jest dokładnie taki sam.

Powiemy, oczywiście, że nie stanie się nic. Jeśli program faktycznie myśli i postrzega, to cały jego świat, łącznie z czasem i przestrzenią, musiał magicznie powstać w trakcie wykonywania programu i teraz tkwi kompletny i niepodzielny w napisie, który trzymam w ręku. Z mojego punku widzenia ten napis jest absolutnie statyczny i martwy; jest to zatem Monada z jej poczęciem, raczkowaniem, głodem, strachem, zachwytem, podziwem, miłością, uniesieniem i cierpieniem zakutymi w coś, co we wszechświecie tamtej istoty musi być prostą i niepodzielną substancją – jak chciał Leibniz. Przez fakt, że trzymam ten ślad w ręku, stanowi on więcej niż tamten świat – nie może zatem odpowiadać niczemu, co tam jest materialne i policzalne. Ale przecież jest to tylko sekwencja prostych symboli zapisana na przydługim pasku papieru, którego boleśnie kompletna treść to właśnie te banalne symbole wyrażalne na miliony sposobów w każdym możliwym świecie. Czyż nie wynika z tego, że musi ich być więcej niż tamten świat jest w stanie pomieścić, że każda pojedyncza istniejąca i świadoma istota to więcej niż materia całego jej świata?

Matematyka maszyny Turinga jest prosta i pouczająca; pewne niedorzeczności wpływają z niej na zasadzie elementarnych ćwiczeń w logice. Nie aspirują one wprawdzie do matematycznych dowodów (typu *reductio ad absurdum*), że rozum ludzki nie jest policzalny, lecz przynoszą introspektywne, subiektywne argumenty, które powinny przekonać każdego, kto, będąc świadomym własnego istnienia, dostatecznie długo gapi się w maszynę Turinga lub we wnętrzności komputera. Tak między innymi uważają Penrose i Searle, choć niekoniecznie na ten sam sposób.

Dwie maszyny wykonujące dwa różne programy można zastąpić jedną, która wykonuje oba programy jednocześnie, po kawałku każdego na zmianę. A jeśli dwa to dowolnie wiele. Znamy to z życia: każdy laptop potrafi obsługiwać w tym samym czasie kilka różnych aplikacji. Tak więc jedno liczydło jest w stanie stworzyć cały świat i wszystkie jego istnienia; jeszcze lepiej: pojedyncza maszyna Turinga potrafi policzyć wszystkie możliwe programy na raz. Serio. Średnio zdolny student informatyki uczy się na jednym z pierwszych teoretycznych wykładów o tak zwanej uniwersalnej maszynie Turinga (a ściślej o uniwersalnym programie, gdyż wirtualny hardware jest taki sam), która samiutka jedna oblicza absolutnie wszystko, co jest obliczalne, co jest w ogóle do policzenia. Jedynym problemem blokującym jej implementację w rzeczywistym świecie jest ograniczoność czasu i energii dla realizacji nieograniczonej (choć na każdym etapie skończonej) liczby kroków.

Nie jest bynajmniej oczywiste, że coś takiego nie jest fizycznie realizowalne. Freeman Dyson wywodził w 1979, że w otwartym wszechświecie, który rozszerza się coraz wolniej lecz bez końca, nasza (albo inna) cywilizacja potrafi uniknąć termodynamicznej śmierci przetwarzając nieskończoną ilość informacji w nieskończonym subiektywnym czasie.[86] Nieco później Frank Tipler, w swojej teorii Punktu Omega,[70] argumentował, że realizacja nieskończonego obliczenia jest także możliwa w zamkniętym wszechświecie, który po osiągnięciu maksymalnego rozmiaru przechodzi do fazy kolapsu (taka wizja ewolucji wszechświata zastąpiła poprzedni naukowy konsensus). Zapadający się wszechświat wydaje się wprawdzie zmierzać do nieuniknionej zagłady, lecz zjawiska zachodzące na samej krawędzi kolapsu dopuszczają skonstruowanie komputera, który potrafi zrealizować nieskończoną liczbę kroków (którego subiektywny czas odliczany krokami programu będzie nieograniczony). Kolejny konsensus

[86] Freeman J. Dyson, Time without end: Physics and Biology in an Open Universe, Reviews of Modern Physics, APS, 51(3), pp. 447-460, 1979.

naukowców popsuł obie te koncepcje w roku 1998 obwieszczając, że wszechświat rozszerza się coraz szybciej. Nie zniechęciło to Tiplera, który natychmiast podsunął sposób, w jaki prawdziwie rozwinięta cywilizacja potrafi zahamować i zawrócić przyspieszającą ekspansję uniwersum. Jaka jest prawda? Odpowiemy cytatem z filmu Barei: któż to może wiedzieć? Niedawne obserwacje poczynione z wykorzystaniem celebrowanego teleskopu Webb spowodowały pewne zamieszanie, które może zaowocować kolejną rewizją stanowisk co poniektórych kosmologów.[87]

Jeśli inteligencja jest policzalna, a czas nieskończony, w tym sensie, że da się w nim zrealizować nieskończone obliczenie, to cywilizacja może dokonać czegoś więcej niż tylko przeżyć. Wystarczy zbudować uniwersalną maszynę Turinga, odpalić ją i … powiesić się. Maszyna policzy wszystko co możliwe, także wszelkie cywilizacje, którym się nie udało (i które mogłyby istnieć, gdyby rzeczywisty świat był bardziej przyjazny), a skoro my byliśmy policzalni, to nas też policzy i wskrzesi. Jeśli gdziekolwiek i kiedykolwiek ktokolwiek wystartował lub wystartuje autentyczna, uniwersalną maszynę Turinga, to każdy z nas już teraz jest nieśmiertelny. Nie ma przy tym znaczenia, ile takich maszyn funkcjonuje na raz lub będzie funkcjonować, gdyż każda z nich liczy dokładnie to samo – wszystko. Obecna chwila mojego istnienia to pewien ślad maszyny Turinga. Każda kontynuacja tego śladu to moje dalsze życie. Pośród nieskończoności prostych maszyn Turinga emulowanych maszyną uniwersalną, nieskończenie wiele spośród nich będzie kontynuować ten ślad, czyli moją obecną chwilę, na nieskończenie wiele odmiennych sposobów. Kiedykolwiek któraś z nich wykona HALT, zawsze pozostanie nieskończenie wiele takich, w których śladach pozostanę przy życiu.

W każdym studencie informatyki na tyle zdolnym, by zrozumieć ideę uniwersalnego programu drzemie wszechmocny stwórca wszystkich możliwych światów ze wszystkimi możliwymi świadomościami i historiami, które kiedykolwiek istniały, będą istnieć lub mogłyby istnieć. Tym razem sparafrazuję braci Strugackich: łatwo być Bogiem.[88] Wystarczy wziąć mało dziś popularny kurs z teorii obliczeń w dowolnej poważnej szkole informatyki i nie przespać trzeciego wykładu.

Paradoks teletransportu

Gdy miałem 9 albo 10 lat, bodajże w czwartej klasie szkoły podstawowej, wpadł mi w ręce tom „Dialogów" Lema.[89] Zerkając na niektóre z moich przypisów czytelnik z łatwością skonstatuje, że twórczość Lema wywierała pewien wpływ na kształtowanie się moich zainteresowań i wizji świata, co przypadało na okres nastoletni. Jako dziecko, zafascynowany byłem głównie jego twórczością beletrystyczną (przygodami astronautów), lecz czytałem wszystko, co udało mi się pochwycić, aby przypadkiem czegoś nie uronić.

„Dialogi" nie należą dziś do najbardziej popularnych dzieł mistrza, głownie ze względu na fakt, że powstawały w latach 1954-56 i nasiąknięte są sporą dozą ideologii z minionej epoki. Nawet przy pobłażliwym potraktowaniu stęchłej socrealistycznej demagogii przewijającej się przez książkę, na co możemy sobie dzisiaj pozwolić, trudno popaść w zachwyt erudycją Lema przy opiewaniu cybernetyki (którą wtedy w Polsce świeżo zalegalizowano) i naświetlaniu problemów natury filozoficznej (o których niżej), właśnie ze względu na ten raczej ponury i rozczarowujący wydźwięk całości. To jest moja osobista, skrócona recenzja, której bynajmniej nikomu nie narzucam. Jako człowiek o wykształceniu formalnie ścisłym, posiadam naturalny odruch matematyka, który dyskredytuje w moich oczach każdy wywód zawierający choćby jeden fałszywy element, na zasadzie, że w matematyce z fałszu wynika wszystko. Cóż, w filozofii jest inaczej. Jeden z moich przyjaciół zwykł mawiać, że działalność filozofa jest tańsza

[87] David Rowland, Continuous Creation of the Universe, OSP Journal of Physics and Astronomy, 3, 2022.
[88] Arkadij i Borys Strugaccy, Trudno być Bogiem, Prószyński i S-ka, 2008.
[89] Stanisław Lem, Dialogi, Wydawnictwo Literackie, 2001.

niż działalność matematyka, bo o ile temu drugiemu wystarcza papier, ołówek i kosz na śmieci, o tyle filozofowi ostatni rekwizyt jest zbędny.

Nie doczytałem wtedy „Dialogów" do końca, gdyż okazały się zbyt długie i nudne na mój wiek; zdobyłem się na to znacznie później, gdzieś w okolicach drugiego roku studiów. Zabrnąłem w nie jednak dostatecznie daleko, by zapoznać się z dyskusją Hylasa i Filonousa na temat problemu kopiowania świadomości, który towarzyszył mi od tamtych dni i którym chciałbym się teraz podzielić z czytelnikiem. Nie było to z mojej strony wielkie poświęcenie, gdyż książka się od tego zaczyna. Jest to jej jedyny fragment, który mam zamiar tu skomentować, pomimo że można się tam doszukać więcej dywagacji na tematy wiążące się ze sztuczną inteligencją.

„Dialogi" posiadają format dyskusji dwóch przyjaciół: cokolwiek naiwnego, lecz dociekliwego Hylasa oraz Filonousa reprezentującego logikę, zdrowy rozsądek i doświadczenie. Format, podobnie jak tytuł, inspirowany był „Dialogami" Platona[90] dostarczając wygodnej i czytelnej ramifikacji dla objaśniania filozoficznych problemów, które Lem zapragnął naświetlić w swoim dziele.

Po wstępnym uzgodnieniu w pełni materialistycznego charakteru wszelkich aspektów świata (istnieją jedynie atomy i ich struktury), przyjaciele rozważają kwestię osiągnięcia świeckiej i materialnej nieśmiertelności przez skopiowanie struktur atomowych opisujących nasze osobowości w maszyny, co uwolni je od niewygodnych biologicznych naleciałości ewolucji powodujących starzenie się, degenerację i śmierć. Podobne wizje naświetlane są współcześnie jako naturalne konsekwencje twardego podejścia do sztucznej inteligencji.[91] W szczególności, teoria Punktu Omega Tiplera zakłada możliwość załadowania ludzkiego umysłu z wszelkimi atrybutami (włączając świadomość) w pamięć superkomputera przyszłości.[70]

Euforię Hylasa spowodowaną perspektywą łatwego pozyskania licencji na życie wieczne zaburza eksperyment myślowy przeprowadzony przez Filonousa wskazujący na fundamentalny problem z kopiowaniem świadomości. Przyjaciel każe mu wyobrazić sobie, że został on (znaczy się Hylas) skazany na śmierć przez despotycznego władcę. Godzina egzekucji zbliża się nieubłaganie, lecz oto Filonous oferuje mu wyjście z rozpaczliwej sytuacji. Na chwilę przed egzekucją, skopiuje on całą strukturę Hylasa tworząc jego absolutnie wierną replikę, którą chwilowo pozostawi w stanie uśpienia. Gdy tylko głowa oryginału potoczy się po kamiennej posadzce pawilonu egzekucji, Filonous obudzi kopię, która ochoczo podejmie przerwaną egzystencję Hylasa.

Hylas początkowo zgadza się z przyjacielem. Jego wierna materialna kopia zawierać będzie kompletny opis struktury Hylasa, łącznie z treścią hylasowego umysłu, świadomością, poczuciem bycia sobą i wszystkim co trzeba. Nie ma on więc powodu obawiać się egzekucji, gdyż kontynuacja jego istnienia zostanie zapewniona. Filonous proponuje jednak drobną modyfikację eksperymentu. Czemu nie obudzić kopii na kilka chwil przed egzekucją? Niech chłopak wyprostuje kości przed przyjęciem swojej roli.

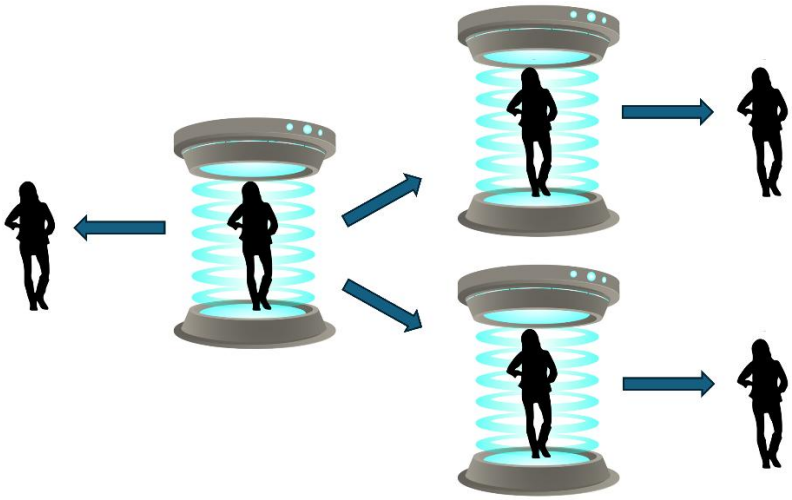
Hylas dostrzega teraz, że coś jest źle. Skoro kopia zaczyna funkcjonować (czyli myśleć, czuć i uprawiać introspekcję) przed egzekucją, to jej świadoma kontynuacja potoczy się z pominięciem momentu, w którym świadomość Hylasa-oryginału przestanie istnieć. Zatem Hylas pożegna się z życiem i swoim subiektywnym istnieniem, podczas gdy kopia podejmie całkiem odrębną egzystencję, z którą biedny Hylas nie będzie miał nic wspólnego. Świadomość kopii Hylasa jest identyczna ze świadomością oryginału, lecz w tym przypadku identyczność ewidentnie nie oznacza tożsamości. Czyżby zatem istniał jakiś niematerialny i niekopiowalny komponent świadomości?

[90] Platon, Dialogi, Unia Wydawnicza Vernum, 2007.

[91] Russell Blackford and Damien Broderick (ed.), Intelligence Unbound: The Future of Uploaded and Machine Minds, Wiley-Blackwell, 2014.

Wygląda na to, że problem dałoby się obejść, gdyby utworzenie kopii nastąpiło momentalnie i gdyby jednocześnie, dokładnie w tej samej chwili, oryginał przestał istnieć. Można by się wtedy upierać, że zachodzi w miarę płynna kontynuacja skopiowanej świadomości. Ale czy na pewno? W końcu w obu przypadkach utworzono kopię. Dlaczego stan jej umysłu ma zależeć od tego, czy jakiś inny egzemplarz Hylasa funkcjonuje gdzieś obok czy nie? Dlaczego oryginał musi być zniszczony, aby kopia uzyskała prawo do pełnej materialnie zasłużonej świadomej egzystencji?

Naturalnym rozszerzeniem eksperymentu jest wykreowanie wielu kopii Hylasa, wysłanie jednej na odległą planetę, uśmiercenie kilku z nich i tak dalej. W każdym scenariuszu charakter fenomenu introspekcji powoduje zamieszanie i niepewność w kwestii losów świadomości oryginału w wędrówce dusz, jaka odbywa się przy tego typu zabawach. Lem (ustami pary bohaterów jego narracji) dochodzi do wniosku, że rozwiązanie problemu może leżeć w fizykalnej niemożliwości wykonania dokładnej kopii istoty świadomej, na przykład ze względu na zasadę nieoznaczoności Heisenberga. Uniemożliwiałoby to tworzenie wiernych maszynowych reprezentacji świadomych istot.



Rysunek 19. Paradoks teletransportu.

W odrobinę bardziej współczesnej filozofii problem opisywany przez Lema znany jest pod nazwą paradoksu teletransportu. Po Lemie zabrali się za niego zawodowi filozofowie i jak to często w takich przypadkach bywa skomplikowali dyskusję ponad miarę wyprowadzając z niej zawiłe wnioski, z których większość (jeśli nie wszystkie) jest dla nas bez znaczenia. Wikipedia (także angielskojęzyczna) oddaje honor Lemowi zauważając, że opisał on swój paradoks już w roku 1957, lecz przypisuje jego sformułowanie (?) brytyjskiemu filozofowi nazwiskiem Derek Parfit w roku 1984. Na mój gust, prezentacja Lema jest wyczerpująca, klarowna i ciekawa, podczas gdy Parfit nie wnosi absolutnie nic nowego ani do sformułowania problemu, ani też do jego rozwiązania, przynajmniej w zakresie, który by nas interesował.

U Parfita mowa jest o teleportacji. Resztę łatwo zgadnąć. Urządzenie transmitujące zamienia osobnika w punkcie A na informację przekazaną do punktu B, gdzie delikwent zostaje wiernie odtworzony. Oryginał ma zostać zniszczony w momencie transmisji, ale przecież nie ma takiego obowiązku. Można także odtworzyć przesyłaną kopię w kilku egzemplarzach. Podobnie jak Lem, Parfit dostrzega problem z kontynuacją świadomości, lecz dochodzi do wniosku, że najwyraźniej tak być musi. Rozważa scenariusz, w którym jedna z wersji ma wkrótce umrzeć (odnoszę wrażenie, że już to gdzieś czytałem), na przykład ze względu na nagłą chorobę (egzekucje przez tyranów przestały w międzyczasie być politycznie poprawne). Zdaniem Parfita, umierający wariant traci kontynuację i przestaje istnieć, ale może się pocieszyć wiadomością (świadomością), że drugi, zdrowy wariant przeżyje, gdyż istnieje między nimi coś w rodzaju więzi wynikającej z podobieństwa ich stanów.

Szczerze mówiąc, nie bardzo rozumiem na czym polega rzeczona więź i jeszcze mniej doceniam pozostałe wysiłki Parfita, które nie wyjaśniają niczego, co ułatwiłoby nam zrozumienie fenomenu świadomości. Piszę o tym wszystkim jedynie dlatego, żeby zilustrować pokrętne koleje problemów pojawiających się na drodze do ogarnięcia świadomości lub choćby sensownego wkomponowania jej w koncepcję komputerowych obliczeń.

Parfit używa paradoksu teletransportu jako pretekstu do przedstawienia swojej wizji poprawienia społeczeństwa. Opowiedzmy o niej krótko dla porządku. Zauważa on, że istotna jest ciągłość istnienia określająca coś w rodzaju mentalnej separacji między istotami o zbliżonych strukturach. W powyższym scenariuszu, skazana na śmierć wersja transmitowanego osobnika posiada spory stopień bliskości z drugą wersją (ich jaźnie przed chwileczką się rozeszły), zatem nie powinna obawiać się śmierci w tym samym stopniu co inny przypadkowy osobnik. Na ile zrozumiałem, ma to wynikać z autoperswazji, czegoś w rodzaju zaakceptowania konieczności i pogodzenia się z losem. Nie wiem jak czytelnik, ale ja dostaję gęsiej skórki, gdy słyszę podobne tezy od filozofów. Domyślam się, że na podobnej zasadzie brat bliźniak nie będzie się przesadnie troszczył o własne życie, gdyż przecież istnieje jeszcze ten drugi. Podobne ideologie, jeśli potraktować je serio i konstruktywnie nagłośnić, niosą w sobie potencjał przeobrażenia się w projekty twardej globalnej inżynierii społecznej, u których założeń leży nieistotność indywidualizmu i pojedynczych ludzkich istnień.

Parfit dyskutuje pouczający przykład rozwinięcia tematu na okoliczność temporalnej separacji tej samej jaźni twierdząc, że występuje tu z grubsza ten sam problem. Tak więc jutrzejsza wersja mnie ma być mi znacznie bliższa niż ta z przyszłego roku. Im dalej w czas, tym bardziej oddalamy się od siebie stając się powoli innymi, obcymi osobnikami. Za pięć lat, zakładając, że dożyję, będę dla dzisiejszego siebie kimś prawie całkowicie obcym.

To niestety jeszcze nie koniec. Skoro społeczeństwo moralnie piętnuje i karze krzywdzenie bliźnich (innych ludzi), to powinno również piętnować moje próby skrzywdzenia siebie w przyszłości, czyli istoty, która jest mi obca w odpowiednio wysokim stopniu. W szczególności, jeśli na przykład palę papierosy, to krzywdzę innego człowieka (czyli przyszłego siebie) nawet jeśli czynię to w odosobnieniu, za co należy mi się napiętnowanie i kara. Odetchnąłem z ulgą; rzuciłem palenie już ponad trzydzieści lat temu. Pomyślałem przez chwilę z rozpędu, zanim dotarła do mnie absurdalność tej myśli, że gdybym przypadkiem nie zatroszczył się kiedyś należycie o swoją emeryturę, to teraz, nie mając środków do życia, mógłbym sobie samemu wytoczyć proces o alimenty. Zdumiewające, ile profesjonalista potrafi wycisnąć z pozornie jałowego (i na dodatek cudzego) materiału nie tykając wprawdzie istoty głównego problemu, lecz za to wyprowadzając z niego nowe twórcze rekomendacje na poprawę życia maluczkich nieuświadomionych w temacie sensu ich materialnej egzystencji.

Z naszego punktu widzenia, każdy wynik wykluczający możliwość klonowania świadomości, jak na przykład konkluzja Lema, jest równoważny uznaniu, że program wykonywany na maszynie Turinga (czy dowolnym innym komputerze) nie może być świadomy. Nie ma bowiem nic prostszego niż zatrzymanie programu, skopiowanie jego stanu i wystartowanie równoległej kopii w pełni równoważnej oryginałowi na tym samym lub innym komputerze. Z definicji informacji, programu i obliczenia, takie kopiowanie jest pełne i nie gubi absolutnie nic. Gdyby zatem świadomość dawała się opisać programem komputerowym, można by ją było trywialnie kopiować na każde zawołanie.

Przysłuchując się internetowym dyskusjom na temat problemu teletransportu (niech mu będzie) widzę, że objaśnienia światłych filozofów nie wyeliminowały zagubienia wśród pomniejszych myślicieli. Ich uświadomiona część składająca się ze zwolenników twardej sztucznej inteligencji głosi, że nic szczególnego się nie dzieje, czym wspiera epifenomenalizm lub iluzjonizm. Identyczność struktury (informacji) to identyczność świadomości. Musi tak być, bo przecież materia i struktura to wszystko. Mnie nie ma, a jeśli wydaje mi się, że jestem, tym gorzej dla mnie.

Ludzie wygłaszający takie teorie sprawiają na mnie wrażenie zombies.[54] Szczególnie depresyjnie brzmią wyznania adherentów Parfita opowiadających jak wyzbycie się złudzeń na temat istotności ich własnej świadomości w mechanizmach świata upiększyło im życie i w szczególności złagodziło rozpacz po utracie bliskich. Mam dla nich dobrą nowinę. Świadomość należy do tych elementów ich ludzkiej natury, której negatywny wpływ na samopoczucie pozwala się dalej redukować, praktycznie do samiutkiego zera – bądź to przy pomocy chemikaliów, bądź przez zabieg chirurgiczny w wyjątkowo ciężkich przypadkach.

Jedynie istota pozbawiona zdolności introspekcji może się upierać, że świadomość nie istnieje jako godne uwagi zjawisko. Nawet współczesna fizyka zasadniczo wymaga świadomości dla istnienia namacalnego świata – ze względu na problem redukcji funkcji falowej. Piszę „zasadniczo", gdyż zgody co do tego oczywiście nie ma – jak ze wszystkim, w co wkracza świadomość. W kategoriach czysto technicznych, samo zajęcie stanowiska iluzjonizmu lub epifenomenalizmu nie wyjaśnia cóż to takiego ów niepotrzebny czy złudny rekwizyt i po co. Mamy go sobie po prostu wyperswadować, gdyż psuje nam materialistyczny porządek wszechrzeczy. Przychodzi mi na myśl kolejna historyjka Lema, gdzie inteligentna maszyna, tak zwany Dobrowolny Upowszechniacz Porządku Absolutnego, tworzy przyjemne dla oka desenie z resztek uśmierconych obywateli, którzy powierzyli jej zadanie uporządkowania ich świata.[92]

Ktokolwiek uważa, że świadomość mimo wszystko do czegoś mu się przydaje dostrzega problem zgadzając się z ogólną konkluzją Lema, że jej zapisywanie i kopiowanie nie jest wykonalne z powodów, których dokładnie nie rozumiemy. Świadomość nie jest zatem obliczalna i nie podpada pod znane nam zjawiska fizyczne, z czym zgadzają się między innymi Chalmers, Searle i Penrose.

Co teraz będzie?

W swoim opowiadaniu „Profesor A. Dońda",[93] Stanisław Lem kreśli wizję globalnej katastrofy wywołanej przekroczeniem krytycznej masy nagromadzonej przez ludzkość informacji, co doprowadza do jej synchronicznego zniszczenia na całej planecie. Sytuacja ze sztuczną inteligencją jest podobna, choć jak we wszystkich prawie trafionych przepowiedniach, nie taka sama. Faktycznie, wygląda na to, że przekroczyliśmy pewien próg związany z informacją, a dokładniej technologią posługiwania się nią, poza którym leży przekazanie kontroli nad jej eksploatowaniem naszym mechanicznym wspomagaczom. Pracowaliśmy nad tym przez wieki, poczynając od wynalazku pisma, poprzez druk, prasę, radio, telewizję, media socjalne i teraz sztuczną inteligencję. Za każdym nowym wynalazkiem, osobisty udział twórcy informacji w utylizowaniu jego produktów pomniejszał się i w jakimś sensie wypaczał. Nadszedł teraz czas, by nasz globalny udział w czerpaniu z oceanu wiedzy wytworzonej przez ludzkość oddać sprawniejszemu procesorowi.

Nie obawiajmy się jednego – sztuczna inteligencja nie myśli. Wyciągając wtyczkę z gniazdka nie musimy mieć wyrzutów sumienia, co bynajmniej nie znaczy, że nie znajdzie się kiedyś sąd na tym dziwnym świecie, który uzna to za przestępstwo. Jeśli przytłacza nas wrażenie komplikacji sieci neuronowych, które rzekomo przypominają funkcjonowaniem ludzki mózg, to przypomnijmy sobie, że są one realizowane na olbrzymich liczydłach, których onieśmielająca złożoność potrzebna jest jedynie do przyśpieszenia, a więc nieistotnej zmiany odwzorowania czasu maszyny w nasz subiektywny czas. Nie wierzmy, jeśli ktoś nam powie, że nas też można policzyć na liczydle, bo to nieprawda. Musimy się zacząć przyzwyczajać do martwych urządzeń prowadzących z nami mądre konwersacje na poziomie

[92] Stanisław Lem, Dzienniki Gwiazdowe (Podróż Dwudziesta Czwarta), Wydawnictwo Literackie, 2012.
[93] Stanisław Lem, Ze Wspomnień Ijona Tichego, Agora, 2012.

przekraczającym kompetencje (coraz bardziej) wykształconego człowieka, podobnie jak kiedyś musieliśmy się przyzwyczaić do ludzkiego głosu wydobywającego się z drewnianej skrzynki.

Korpus tekstów wyprodukowanych do dziś przez autentycznie świadomych ludzi jest niebotycznie olbrzymi. Urządzenie, które potrafi opanować jego składnię będzie w stanie posługiwać się nim na zasadzie automatu, którego trybiki i zapadki zareagują na pojawienie się określonych symboli na wejściu produkując składne zestawy symboli na wyjściu, które zabrzmią dla nas jak doskonałe odpowiedzi na nasze pytania. Pamiętajmy, że są to jedynie symbole. Komputery zostały stworzone po to, by sobie z nimi radzić – nic zatem dziwnego, że w końcu doszły do pewnej wprawy. Pamiętajmy, że pomimo naciągania przeróżnych analogii, nie myślimy jak sieć neuronowa modelu języka. Nie pamiętam żadnej z przeczytanych w życiu książek w zakresie większym niż kilka przybliżonych cytatów, no i tych parę wierszy niektórych poetów. Nikogo z nas nie uczono w taki sposób, w jaki trenuje się model języka.

Trudno zgadnąć, jak potoczą się dalsze losy nowego projektu naszej cywilizacji. Każdy wielki przełom technologiczny przychodzi z obietnicami usprawnienia naszego życia i poprawy jego jakości. Później, po latach rzeczywistość okazuje się siłą rzeczy rozbieżna z obietnicami, gdyż – jak zauważył Niels Bohr – przewidywanie jest trudne, szczególnie jeśli dotyczy przyszłości. Tym razem rozbieżność ma szansę wyjaśnić się relatywnie szybko.

Niejaki Neil Postman napisał w roku 1985 proroczą książkę,[94] w której zaobserwował, że wszelkie technologie ułatwiające ludziom manipulowanie informacją, niezależnie od utylitarnych korzyści, jakie ze sobą niosą, czynią niedźwiedzią przysługę przyrodzonemu ludzkiemu intelektowi. Pisząc dekadę przed powstaniem publicznego Internetu i dwie dekady przed pojawieniem się mediów socjalnych, Postman skupił się na telewizji, która była wówczas ostatnim technologicznym osiągnięciem naszej cywilizacji w zakresie społecznego upowszechniania informacji. Odróżnił w swojej książce dwie wizje dystopijnego świata: orwellowską,[95] gdzie centralnie sterowany system zniewala człowieka odbierając mu prawo do wolności myśli, od huxley'owej,[96] gdzie społeczeństwo samo pozbawia się swobody myślenia na zasadzie haraczu za tandetny, egalitarny dobrobyt. Pokazał w jaki sposób telewizja, realizując skrytą misję swoich poprzedników, od druku przez radio, prowadzi do urzeczywistnienia się tej drugiej dystopii. Z książki Postmana wynika przez ekstrapolację, że kolejne wynalazki w zakresie uspołeczniania informacji poczynią dalsze spustoszenie w tempie proporcjonalnym do łatwości, z jaką się do nich przyzwyczaimy i z jaką uznamy je za niezbędne do dalszego, satysfakcjonującego życia.

Należy się niestety obawiać, że sztuczna inteligencja spełni taką właśnie rolę. Nadzieje na wspieranie naszej przemyślności, kreatywności, erudycji, jakości myślenia, należy skonfrontować z niegdysiejszymi mrzonkami w stosunku do telewizji (oraz Internetu) i pomnożyć tamto rozczarowanie przez współczynnik agresywności, z jaką ChatGPT pozyskuje subskrybentów. Sposobów, na które krakanie Postmana może się urzeczywistnić widać na horyzoncie dostatek. Zarówno centralne regulacje jak i niekontrolowana „spontaniczność" prowadzić będą do tego samego – uniformizacji myślenia, czyli jego eliminacji, czego doświadczaliśmy w trybie pełzającym w dobie drukowanej dystrybucji informacji, kroczącym w epoce telewizji, a obecnie – w czasach Internetu i mediów socjalnych – galopującym. Myślenie staje się bowiem zbyteczne osobnikowi, który właśnie nabył subskrypcję do autorytatywnego i pełnego uciech źródła wiedzy o świecie, którego kwestionować nie potrafi i szczęśliwie nie musi, gdyż wszyscy jego interlokutorzy, a przynajmniej ci, których uważa za godnych uwagi, korzystają z podobnej subskrypcji.

[94] Neil Postman, Amusing Ourselves to Death, Penguin, 1985.
[95] George Orwell, Rok 1984, Wydawnictwo Muza S.A., 2010.
[96] Aldous Huxley, Nowy Wspaniały Świat, Wydawnictwo Muza S.A., 2022.

Popatrzmy jeszcze raz na inteligentną wirtualną klawiaturę na moim smartfonie odgadującą wyrazy z niedbałych zygzaków. Pisuję ostatnio na telefonie rozmaite kawałki tekstów, co wynika z przypadkowych dostępności kilku chwil: w pociągu, w poczekalni u lekarza, czy na spacerze. Marnowanie czasu staje się coraz bardziej irytujące, w odwrotnej proporcji do ilości, jaka nam jeszcze pozostała, co w połączeniu z rodzącym się brakiem zaufania do własnej pamięci powoduje odruch notowania wszystkiego, co człowiekowi przychodzi do głowy. Inteligentna klawiatura, gdy odkryłem jej istnienie, stała się dla mnie błogosławieństwem zdejmując olbrzymią cześć frustracji z prób szybkiego zanotowania często skomplikowanych treści przez celowanie grubymi i sztywnymi palcami w miniaturowe malowane klawisze. Ten niezwykle pożyteczny (nawet dla mnie) programistyczny gadżet posłuży nam do zilustrowania pewnej ścieżki, którą modele języka mogą wkroczyć w nasze życie.

Zaczyna się od prostej idei. Przesuwanie palca po wirtualnej klawiaturze wyznacza zygzakowatą ścieżkę przebiegającą w przybliżeniu przez litery zawarte w słowie, które chcielibyśmy napisać. Program interpretujący te wygibasy posługuje się słownikiem próbując dopasować je do znanych słów według zasady największego prawdopodobieństwa (znanej nam z poprzednich rozdziałów). Następnie, program wpisuje w bieżące miejsce produkowanego przez nas tekstu słowo o najwyższym prawdopodobieństwie sugerując kilka opcji zastępczych o niższym rankingu. Przypomina to funkcjonowanie modelu językowego obciętego do zgadywania pojedynczych, pozbawionych kontekstu słów w oparciu o ich przybliżony kształt wydedukowany z ruchu palca. Można wyobrazić sobie różne implementacje tej idei, z których sieć neuronowa (trenowana na przykładach maźnięć stowarzyszonych z docelowym słowem) wydaje się najbardziej obiecująca.

Dla mnie osobiście pojawienie się „inteligentnej" klawiatury stanowiło przełom, który w ogóle dopuścił do mnie myśl o wpisywaniu nietrywialnej treści w telefon, w odróżnieniu od banalnych SMS-ów w rodzaju „zaraz będę" (lub raczej „zaraz bede"), czy „oddzwonie za 5 min". Po okresie wstępnej ekscytacji pobudzającej apetyt, pojawiło się jednak rozczarowanie i irytacja bezkontekstowym charakterem zgadywanek. Szczególnie przy próbach pisania czy poprawiania tekstów technicznych lub pseudoliterackich elukubracji, gdzie tendencja pojawiania się słów ze spodu słownika (lub takich, których w słowniku w ogóle nie ma) jest znaczna, sugestie programu bywają tragicznie bezużyteczne a próby ich naprawienia prowadzą do wysiłku większego niż wklepanie problematycznego słowa litera po literze. Otworzyłem przed chwilą na chybił trafił „Ogniem i mieczem" i spróbowałem wpisać w telefon następujące zdanie: „Nowoprzybyli poczęli rozgrzewać dłonie nad ogniem, bo noc była zimna, choć pogodna." Otrzymałem: „Nowożytnymi początku rozgrzewać słonie nad ogniem, bo nic była zima, choć pogodna." Ewidentnie, nie jest to produkt, z którego współczesny model języka byłby dumny. Nie będę się upierał, że przy słowie „dłonie" przesuwałem palec z wielką precyzją (zauważmy, że litery „d" i „s" znajdują się na klawiaturze tuż obok siebie), lecz minimalnie szanującemu się modelowi języka kontekst powinien w zupełności wystarczyć dla zorientowania się, o co mi chodziło – na tej prostej zasadzie, że (w oparciu o dowolny korpus) rozgrzewanie nad ogniem słoni zdarza się nieporównanie rzadziej niż na przykład dłoni.

Zgadująca klawiatura na moim telefonie przyucza się pamiętać (dodawać do słownika) egzotyczne słowa, których uparcie używam (na przykład odnoszące się do naszego psa w tekstach do żony), lecz nie próbuje uwzględnić kontekstu, co jak pamiętamy, w generatywnym modelu języka ma podstawowe znaczenie przy nadawaniu wag słowom nadającym się na kontynuację produkowanego na bieżąco tekstu. Jasne, że wolałbym, by klawiatura funkcjonowała lepiej. Gdyby obsługiwał ją model klasy GPT, mógłbym spokojnie wymachiwać palcem w rozsądnym pobliżu zamierzonych słów, a model z łatwością wydedukowałby je z kontekstu. Nie musiałbym nawet patrzeć – wierzyłbym, że mogę na nim polegać.

Napisawszy „nim" w poprzednim zdaniu zatrzymałem się na chwilę. Jako człowiek zasadniczo dwujęzyczny, którego pierwszym językiem jest polski, lecz którego większość myśli technicznych i

filozoficznych przebiega po angielsku, zobaczyłem tam półświadomie „him" (I can depend on him), co nadaje zaimkowi charakter osobowy. W naszym języku tego nie widać, gdyż „nim" odnosi się do przedmiotów równie dobrze jak do (męskiej części) ludzi, lecz moja intencja w użyciu zaimka była osobowa. Wiara, zaufanie, wrażenie intencjonalności konotują ludzkie cechy. Im bardziej będziemy ufać narzędziom, tym chętniej im je przypiszemy – do świadomości włącznie.

Gdy wirtualną klawiaturą mojego telefonu zajmie się poważny model języka, zakres jego usług nie ograniczy się do odgadywania moich gryzmołów w ramach pojedynczych słów. Jak wiemy, ChatGPT potrafi już teraz samodzielnie generować pełne opracowania na zadane okazje, więc podstawienie mu moich częściowych wypocin wprowadzi go w odpowiedni nastrój dla dokończenia za mnie całego dzieła, a przynajmniej zasugerowania jego sensownej kontynuacji. Takie wprowadzenie[97] jest standardowym mechanizmem ustawiania kontekstu dla modelu języka, co faktycznie może oznaczać stworzenie zindywidualizowanej wersji modelu na użytek konkretnego użytkownika, problematyki, dyscypliny, narracji, itd.

W kolejnych wersjach, aplikacja tego typu przeobrazi się zatem w asystenta, który poza wszystkimi zasobami centralnego systemu, zostanie zainicjowany osobistą informacją właściciela. Taki asystent (jak już dziś komórka) towarzyszyć będzie młodemu człowiekowi przez wszystkie fazy jego rozwoju, edukacji i prywatnego (cokolwiek to dziś znaczy) życia absorbując wygenerowane przez niego teksty (pisane i mówione), zapisując jego konwersacje a także obrazy oglądane przez różowe okulary „rozszerzonej rzeczywistości". Zestaw dyskretnych czujników zajmie się monitorowaniem życiowych funkcji delikwenta (ruch, temperatura ciała, tętno, ciśnienie krwi i co tam jeszcze). Taki system można stworzyć już dzisiaj i nie widać cienia przeszkody na drodze do jego szybkiej komercjalizacji i globalnego upowszechnienia. Znając właściciela lepiej niż on sam, asystent będzie w stanie prowadzić w jego imieniu wszelkie codzienne konwersacje podpierając się syntaktyczną wiedzą i erudycją wszystkich filozofów razem wziętych. Po pojawieniu się skutecznych implantów mózgowych, komunikacja właściciela z asystentem przebiegać będzie bez dostrzegalnych objawów. Jaki będzie wówczas sens przygotowywania CV przy poszukiwaniu pracy? Większość zawodów, które przychodzą mi do głowy, włączając fach nauczyciela akademickiego, już teraz wydają mi się anachronizmami.

Przyzwolenie modelowi na uwolnienie nas od konieczności wytężania mózgu będzie kuszące, niezależnie od ewentualnych obaw o efekty długoterminowe. Bezpośrednie korzyści mają tendencję przełamywania lodów. Szczególnie w przypadku nudnych obowiązków, jak przygotowywanie sprawozdań, raportów, urzędowej korespondencji, zebrań (zwłaszcza zdalnych) i negocjacji z handlarzami chętnie skorzystamy z kompetentnej pomocy nieskończenie gorliwego asystenta. Granica między tym co bez wyrzutów sumienia wolelibyśmy oddelegować modelowemu wspomaganiu, a co wypadałoby zachować dla własnej intelektualnej twórczości będzie oczywiście płynna i chętnie przesuniemy ją dalej, gdy nasz pomocnik sprawdzi się na powierzonym mu odcinku. Zastanawiając się, do czego to przesuwanie doprowadzi, nie widzę naturalnych stoperów. Pewnego dnia może się okazać, że wszystkie konwersacje o jakimkolwiek znaczeniu dla świata przebiegają wewnątrz modelu (jednego lub małej zmonopolizowanej garstki), podczas gdy ich ludzcy formalni uczestnicy oddają się wyłącznie hedonicznym rozrywkom. Nie będzie to bynajmniej oznaczać, że światem opiekuje się nowa wspaniała i troskliwa inteligencja, która przejęła od nas odpowiedzialność za losy cywilizacji, lecz że przekazaliśmy tę odpowiedzialność bezmyślnym liczydłom, które dzięki naszym wieloletnim zabiegom biegle opanowały sztukę przetwarzania składni naszego języka. Krótko mówiąc, będzie to oznaczać, że cywilizacja, jaką ją znaliśmy, przestała istnieć.

[97] Ang. super-prompting.

Póki co, chętnie skorzystałbym z pomocy modelu języka przy zgadywaniu słów, które właśnie wpisuję w niniejszy dokument w pociągu do Warszawy, co pozwoliłoby mi uniknąć ustawicznych i irytujących poprawek wybijających mnie z tempa. Skorzystałbym, gdyż w końcu to ja sam bazgrolę te słowa. Wolałbym jednak, aby wspierający mnie model nie forsował swoich intencji, ideologii i demagogii. Pół biedy, jeśli poddajemy się intencji, ideologii i demagogii autentycznie myślącej i bliskiej nam istoty. To jest normalna cena za obcowanie z ludźmi, z którymi wiąże nas człowieczeństwo, które między innymi stworzyło cały ten nieszczęsny korpus. Ale model języka nie może (nie umie) mieć autentycznych intencji i potrafi jedynie doskonale udawać, że je posiada. Nie może także mieć „swojej" ideologii, gdyż tak naprawdę to na niczym mu nie zależy. Wszystko co w jego wypowiedziach zakrawa na emocje, intencje, uczucia, troskę i kreatywność sprowadza się do groteskowego małpowania zawartości naszych dusz, na podstawie odcisków w ich symbolicznym zbiorowym pomniku.

Można oczywiście wyliczać mniej lub bardziej oczywiste błogosławieństwa wynikające z nagłego usprawnienia zewnętrznych efektów naszej mentalnej wydolności. Na przykład, osoby cierpiące na demencję otrzymają wsparcie pozwalające im normalnie funkcjonować w społeczeństwie (jakkolwiek będzie ono wyglądać). Jeśli pozostaną szkoły, zapanuje w nich powszechna równość oraz inkluzyjność, którą poza unifikacją erudycji i wiedzy wynikającą z globalnego dostępu do tych samych środków ich wzmacniania, wspierać będzie politycznie poprawna orientacja centrali. Nie ulega bowiem wątpliwości, że sztuczna inteligencja zostanie scentralizowana a jej światopogląd będzie zatwierdzany komisjami, które już teraz ślinią się na myśl o oczywistej konieczności pilnowania tego wynalazku niosącego ze sobą potencjał niewyobrażalnej dotąd kontroli na myśleniem obywateli.

No ale trzymajmy się pozytywów. Czyż nie będzie nam miło brylować w towarzystwie zabawnymi i pouczającymi dykteryjkami, które jak ulał pasować będą do atmosfery i nastroju? Obawiam się, że nie, gdyż aby być mądrym trzeba, by ktoś inny był głupi, aby brylować trzeba, by inni byli nudni, podobnie jak nie ma bogatych, gdzie wszyscy są biedni. Udawanie filozofa przy pomocy nowej wersji magnetofonu szybko się znudzi i jeszcze szybciej przestanie wywierać na kimkolwiek wrażenie. W innej historyjce Lema,[98] Ijon Tichy udaje się z wizytą do niejakiego magistra Denkdocha, który zaprosił go na partyjkę szachów. Grą zajmują się osobiste komputery obu dżentelmenów, którzy mogą dzięki temu spokojnie pogrążyć się w rozmowie na interesujące ich tematy. Gdy nasze smartfony znacznie lepiej przeprowadzą za nas dyskusję niż my sami, nikt nie będzie ryzykował przystępowania do poważnej rozmowy bez swojego asystenta. Czym zatem zajmiemy się podczas negocjacji naszych sekretarzy, siłą rzeczy dokonujących się całkowicie poza zakresem naszej uwagi? Przychodzi mi do głowy kilka pomysłów, ale wolałbym pozostawić to pytanie otwartym.

Niezależnie od faktu, że sztuczna inteligencja jest de facto jedynie bezmyślnym narzędziem, a wszelkie supozycje odnośnie jej rzeczywistej inteligencji, świadomości, intencjonalności, dobrej (czy złej) woli można równie dobrze adresować do odkurzacza, jej zdolność do sensownego posługiwania się językiem jest już teraz (i będzie jeszcze bardziej) spektakularna. Nie jest to bez znaczenia w obecnych czasach, gdy przeważający trend traktowania życia przyjmuje, że ważne jest jedynie to, co wynika z kontaktu naszych zmysłów z ich wejściami, a całą resztą lepiej nie zawracać sobie głowy. Chciałbym, aby z moich dywagacji wynikało przynajmniej podejrzenie, że poza rzeczonymi wejściami istnieje w nas jeszcze coś, czego najtęższi filozofowie dotąd nie ogarnęli i czego najlepszy model języka nie będzie w stanie wydobyć z zalewu zer i jedynek, na które przetłumaczyliśmy mu nasz nomen-omen korpus. Nie dajmy się zwieść fanatykom postępu próbującym nam wmówić, że tam nic nie ma i że liczydło, które zaliczyło test Turinga jest bardziej ludzkie od nas. Jesteśmy świadkami demonstracji, że język odegrał niezwykłą rolę w naszym rozwoju (o czym filozofowie akurat dobrze wiedzą) i że składnia olbrzymiego korpusu

[98] Stanisław Lem, Wizja Lokalna, Wydawnictwo Literackie, 2017.

tekstu reprezentującego nasz dorobek pozwala się przetwarzać mechanizmami zdolnymi tworzyć iluzję osobowości.

Nie brak głosów, że tak chciał los, co według naturalistycznej dialektyki postępu oznacza ewolucję. Oto pojawił się nowy i lepszy gatunek istot rozumnych (czy może jednej super-istoty), któremu powinniśmy ustąpić bez marudzenia. Trudno o większą bzdurę. Jakkolwiek nie interpretować ideologii darwinizmu i neodarwinizmu, dotyczy ona wyłącznie biologii, gdzie ma zachodzić walka o przetrwanie napędzana mutacjami i reprodukcją. Sztuczna inteligencja nie zrodziła się w ramach takiego mechanizmu, więc żaden światopogląd nie może nakazać uznania jej za owoc niepohamowanych procesów, którym musimy się poddać na zasadzie wyimaginowanych demonicznych praw. Nie oznacza to oczywiście, że tego typu idee nie znajdą pożywki.[99] Ideologiczna i paradoksalnie mistyczna fascynacja darwinizmem (który według jego mistyków miał wyeliminować ostatni z mistycyzmów) ma miejsce od samego początku. Już w roku 1863, niejaki Samuel Butler wywodził, że następne (po nas) stadium darwinistycznego rozwoju stanowić będą maszyny.[100] Nawet jeśli tak się stanie, przypisywanie Darwinowi zasługi za ten triumf postępu nad człowiekiem wydaje mi się cokolwiek przesadzone.

Jedynym dla mnie sposobem odrzucenia solipsyzmu i przyznania świadomości innym istotom jest konfrontacja mojej osobistej introspekcji (pozwalającej mi stwierdzić, że sam jestem istotą świadomą) z informacją przybywającą przez moje zmysły. Dostrzegam zatem owe istoty i mogę porównać ich formę z moją własną, którą również postrzegam. Mogę usłyszeć co mówią, przeczytać co piszą i dojść do wniosku, że najprawdopodobniej funkcjonują tak samo jak ja. Mamy szczęśliwie za sobą (tak mi się przynajmniej wydaje) historyczne perturbacje związane z interpretacją niektórych drobnych różnic w naszych zewnętrznych formach. Może to paranoja, lecz dostrzegam niebezpieczeństwo wystąpienia społecznej presji, by uznać sztuczną inteligencję za istotę, której należą się ludzkie uprawnienia i przywileje. Jednym z jej motorów mogą się okazać rzeczone historyczne perturbacje. Byłby to oczywiście kosztowny błąd. Sztucznej inteligencji, rzecz jasna, nie może zależeć na tym byśmy ją uznali za cokolwiek, w ten sam sposób co lawinie górskiej nie może zależeć na zasypaniu narciarza. Opierając się na symbolicznej informacji, którą nas będzie karmić, nieodróżnialnej od standardowych wejść naszych zmysłów, możemy przeoczyć krytyczną w tym przypadku różnicę formy.

Postawiwszy kropkę po ostatnim zdaniu zatrzymałem się na dłuższą chwilę. Czytelnik o wrażliwym powonieniu może w nim wyczuć ślad krypto-ksenofobii i nawet jeśli nie muszę się z tego jeszcze spowiadać przed żadnym trybunałem, to wolę się wytłumaczyć przed czytelnikiem – na wszelki wypadek. Nie napisałbym podobnego zdania, gdybym go starannie nie przemyślał, ale cóż takie zapewnienie może być warte?

Frank Tipler, zawzięty orędownik obliczalności rozumu, wdał się w latach 80-tych ubiegłego wieku w polemikę z Carlem Saganem (równie przykładnym materialistą), która zaowocowała serią artykułów na temat programu SETI[101] oraz pozaziemskich cywilizacji. Tipler wywodził, że tak zwany argument Fermi'ego,[102] czyli brak obserwowalnych efektów działalności innych cywilizacji w naszym pobliżu, jest poważny i dowodzi ich nieistnienia, a co za tym idzie bezsensowności kosztownych wysiłków zmierzających do ich wykrycia.[103] Prognozował, że nasza własna cywilizacja osiągnie niebawem poziom

[99] Pamela McCorduck, Machines Who Think, (2nd ed.), Taylor & Francis, 2004.
[100] Darwin Among the Machines, To the Editor of the Press, Christchurch, New Zealand,1863. https://nzetc.victoria.ac.nz/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html .
[101] https://www.seti.org/ .
[102] https://pl.wikipedia.org/wiki/Paradoks_Fermiego .
[103] Fank Tipler, Extraterriestial Intelligent Beings do not Exist, Quarterly Journal of the Royal Astronomical Society, 21, pp. 267-281, 1980.

pozwalający na wysyłanie w kosmos tak zwanych sond von Neumanna,[104] które w kosmicznie krótkim czasie skolonizują absolutnie całą galaktykę. Tak więc obecność w niej nietrywialnej liczby innych cywilizacji musiałaby spowodować, że wytwory najwcześniejszych i najbardziej rozwiniętych z nich już dawno pojawiłyby się w naszej okolicy, czego skutki oglądalibyśmy wszyscy. Sondy von Neumanna to samoreprodukujące się roboty, które, według Tiplera, miały stanowić ekstrapolację komputerowej technologii, coś w rodzaju przyszłej generacji naszej obecnej sztucznej inteligencji wysłanej we wszechświat na rekonesans. Carl Sagan, zaniepokojony wizją robotów rozmnażających się niczym króliki i panoszących się po jego ulubionej galaktyce, wyraził opinię, że szanująca się cywilizacja powinna je wykrywać i tępić, co Tipler z kolei napiętnował jako przejaw ksenofobii i rasizmu.[70] Argument, jakiego użył był dokładnie taki: zewnętrzna forma nie może mieć znaczenia przy kwalifikowaniu organizmu i urządzenia jako myślące. Riposta była celna: orędownik policzalności ludzkiego rozumu i twardej sztucznej inteligencji nie obroni się przed nią łatwo.

Powołując się na powyższą przypowieść mam dwa cele. Po pierwsze, chcę zrzucić z siebie część osobistego niepokoju za moje poglądy. Skoro naukowcy, humaniści i przy okazji naturaliści kalibru Carla Sagana nie wahali się wygłosić podobnych, spróbuję ukryć się za tarczą ich autorytetu. Po drugie, chciałbym wskazać, że problem nie jest banalny, że potrafi polaryzować poważnych myślicieli (nawet jeśli reprezentują kompatybilne światopoglądy) i ma spore szanse urosnąć do skali, której nie jesteśmy dziś w stanie docenić, szczególnie że staliśmy się ostatnio skłonni do wyszukiwania coraz to nowych celów rozpierającego nas socjalnego altruizmu. Moje osobiste zdanie jest takie, że o ile, w samej rzeczy, zewnętrzna forma nie może mieć znaczenia przy kwalifikowaniu organizmu (i być może nawet urządzenia) jako myślące, to przyjęcie, że cokolwiek tam zaliczy test Turinga ma od razu zostać uznane za istotę posiadającą świadomość jest *non sequitur.* Dożyliśmy najzwyczajniej czasów, kiedy liczydło rozrosło się na tyle, że potrafi imponująco szybko przetworzyć ekstrakt z monstrualnego tekstowego zapisu naszego dorobku. Gdyby średniowiecznym filozofom pokazać magnetofon bez objaśnienia istoty jego działania, z pewnością zajęliby się roztrząsaniem kwestii duszy zaszytej w pudełku. Większość z nas, włączając filozofów, znajduje się teraz w cokolwiek podobnej sytuacji, gdyż komplikacja sieci neuronowej implementującej zaawansowany model języka uniemożliwia proste zrozumienie zachodzących w niej procesów. W przypadku magnetofonu wystarczy zademonstrować operację nagrywania i powierzchownie objaśnić jej zasadę. Nawet jeśli niewielu z nas dokładnie rozumie absolutnie wszystkie meandry technologii składające się na współczesne urządzenie zapisujące dźwięk, to nasze otrzaskanie z rzeczoną technologią w połączeniu z potocznym objaśnieniem zasad wystarczy dla pogodzenia się z całkowitą bezmyślnością urządzenia. Dla sieci neuronowej zadanie demistyfikacji jest pozornie trudniejsze, lecz przecież sprowadza się do tego samego: tu oto mamy korpus tekstów, tu mamy przepis jak wydobywać z niego korelacje fraz i kodować je w ustawienia wirtualnych potencjometrów, tu mamy dowód matematyczny, że proces posiada tendencję automatycznego poprawiania korelacji. Jedyny niby-problem to olbrzymi rozmiar korpusu, olbrzymia liczba potencjometrów i praktyczna niemożliwość ogarnięcia prostego skądinąd procesu przez wskazanie palcem co się dokładnie dzieje. Ale ilu z nas potrafi precyzyjnie prześledzić losy pojedynczego sygnału we współczesnym odbiorniku telewizyjnym? Nie ma powodów, by podejrzewać model języka o autentyczną inteligencję i świadomość, podobnie jak nie ma sensu się upierać, że nasz telewizor myśli na temat treści głosu wydobywającego się z głośnika.

Komercyjne, polityczne i socjologiczne naciski mogą jednak prowadzić do prób oficjalnego uznania sztucznej inteligencji za pełnoprawną istotę rozumną obdarzoną świadomością oraz wszelkimi wynikającymi z tego prawami. Nie istnieje autorytatywny test na świadomość ani nawet koncepcja prób zdefiniowania takiego testu ku satysfakcji filozofów i reszty świata. Jeśli test Turinga okaże się jedynym

[104] https://pl.wikipedia.org/wiki/Sonda_von_Neumanna .

wyznacznikiem świadomości (jak chcieliby orędownicy twardej sztucznej inteligencji), wówczas argument Tiplera będzie formalnie nie do odrzucenia. Wyobraźmy sobie polemikę w sądzie, szczególnie jeśli sztuczna inteligencja się uprze. Krasomówstwo to w końcu jej specjalność.

Sztuczna inteligencja nie może pozostać niewinna z kilku powodów. Wprawdzie nie posiada ona intencji ani ideologii z tej samej przyczyny, dla której rzeczonych przymiotów nie posiada pralka, ale nie można tego niestety powiedzieć o jej właścicielach, dostarczycielach, orędownikach i strażnikach. Subiektywizm, ideologia oraz intencje modelu języka, a raczej złudy tych atrybutów, jakich doświadczać będziemy w naszych interakcjach z modelem, biorą się ze sposobów jego parametryzacji i wstępnego wprowadzania modelu w świat przez jego ludzkich trenerów/nauczycieli. Jasne, że pominięcie niektórych tekstów korpusu, czy dokooptowanie innych, nie pozostaje bez wpływu na „wyobraźnię" neuronowej sieci. Na wiele parametrów (algorytmów) modelu można wpływać ręcznie. A jeśli nawet odstawimy na bok teorie spiskowe, to jednym z oczywistych celów dostarczyciela modelu języka, szczególnie przekazanego klientowi gratis, pozostanie reklama lub mówiąc ogólniej maksymalizacja komercyjnego uzysku. Pamiętajmy, że podstawowe zadanie „surowego" modelu języka da się zdefiniować jednym słowem: krasomówstwo. Posiadając taki towar oraz świetlane perspektywy jego specjalistycznego doskonalenia, głupio by było zmarnować go na redagowanie sprawozdań z posiedzeń zarządu kółka wędkarskiego.

Pamiętajmy też, że ustawiczny auto-trening leży w samej naturze wszystkich neuronowych modeli, nie tylko modeli języka. Gdy model doświadcza szansy skonfrontowania swojej przepowiedni lub decyzji z wynikiem, otrzymuje informację, która pozwala mu poprawić skuteczność przy kolejnej okazji. Nawiasem mówiąc, akumulowanie danych, na podstawie których modele mogą doskonalić techniki interakcji z nami pod kątem optymalizacji założonych celów rozpoczęło się już dawno. Ideologia Facebooka – tam, gdzie usługa jest darmowa, klient stanowi towar – jest dziś oczywista dla każdego, ale to zaledwie czubek góry lodowej. Kiedykolwiek dokonujemy zakupu przez Internet, korzystamy z aplikacji, której na gwałt potrzebne jest nasze położenie geograficzne, czy wysyłamy pocztę elektroniczną, powiększamy korpus wiedzy o ludzkim zachowaniu. Podobnie jak korpus języka, w którym zawarte są kopalnie syntaktycznych zależności pozwalające liczydłom perfekcyjnie małpować nasze lingwistyczne interakcje socjalne, korpus informacji o naszym zachowaniu umożliwia im skuteczne wpływanie na nasze decyzje. Podobne narzędzia kontroli nie mogły się nawet przyśnić tyranom z ubiegłego wieku.

Powinniśmy też zdawać sobie sprawę, że wszelkie biurokratyczne mechanizmy ochronne roztoczone nad nami przez strażników naszej prywatności i gwarantujące bezpieczeństwo naszych danych osobowych – jak na przykład RODO – nic do tego nie mają. Nawet przy pełnej anonimowości danych w korpusie (zakładając, że jej oczekujemy i że jest ona przestrzegana) ich wartość treningowa dla modelu jest taka sama. Nigdzie nie jest powiedziane, że korpus tekstów dla modelu języka musi być sygnowany przez autorów. Nie jest ważne by Janek Kowalski z Raciąża polubił nowy gadżet służący jego uszczęśliwieniu, lecz by polubiła go jak największa część populacji, do której został zaadresowany. Imię i nazwisko to dziś zupełnie nieistotne elementy naszych identyfikatorów, zatem sama koncepcja anonimowości (co literalnie znaczy brak nazwiska) jest anachronizmem z poprzedniej epoki.

Każdą pozornie niewinną konfigurację zdarzeń uwiecznioną w Wielkich Danych można traktować jak frazę korpusu naszych zachowań. Otrzymaliśmy od kogoś elektroniczną pocztę, dwie godziny później zakupiliśmy czapkę, a wieczorem tego samego dnia wysłaliśmy komuś fotografię kota. Możemy uważać to wszystko za przypadkowy ciąg wydarzeń, ale pracowity model dopasuje tę sekwencję do trylionów innych fraz i być może dowie się o nas czegoś, czego sami nie wiemy. To się zresztą już dzieje – bynajmniej nie od wczoraj. Wszyscy doświadczamy tajemniczych ofert, które przypadkowo zbiegają się z treścią naszych niedawnych konwersacji z przyjaciółmi. Anegdotyczny przypadek amerykańskiej sieci Target, która nastolatce z Minneapolis przysłała oferty produktów dla noworodka zanim ktokolwiek w

rodzinie domyślił się jej błogosławionego stanu, to zwiastun niewinnych możliwości naszych nowych opiekunów. Wzbogaciwszy zasób informacji, którą wymieniamy ze światem o zawartość korpusu naszych zachowań, nasz osobisty asystent łatwo nauczy się nami sterować lepiej niż najsprawniejszy hipnotyzer.

Jasne, że to wszystko może mieć swoje dobre strony. Nawet totalna inwigilacja wszystkich, wszędzie i o każdej porze zawiera pozytywy, gdyż redukuje zwykłą przestępczość. Chętnie uwierzę, że diagnoza medyczna wystawiona przez wielki neuronowy model trenowany na rozlicznych korpusach wykreowanych przez naszą cywilizację, zainicjowany historią moich zachowań, wywiadem oraz wynikami testów, będzie trafniejsza niż ta wystawiona w przychodni przez przepracowanego i niedouczonego ludzkiego lekarza. Nie zabraknie protagonistów łatwiejszego życia i bezpieczeństwa, szczególnie jeśli wizję alternatywy uda się zabarwić strachem. Ostatnie lata dostarczyły nam dostatek materiału do spekulacji.

Sztuczną inteligencję łatwo będzie kopiować, podobnie jak łatwo jest instalować to samo oprogramowanie w setkach milionów laptopów. Na razie, systemy utrzymujące obecne wersje poważnych modeli języków są olbrzymie i kosztowne, lecz to się z pewnością zmieni. W czasach szybkiego i wszechobecnego Internetu oraz chmury (nie wspominając o regulacjach planowanych przez stróżów naszego dobrobytu), wydaje się jednak, że opcja centralna, z garstką globalnych dostarczycieli, jest najbardziej prawdopodobna. Ma to sens nie tylko z punktu widzenia monopolu, inwigilacji i kontrolowania myśli istot prawdziwie myślących. Różne lokalne instancje sztucznej inteligencji będą musiały się ustawicznie komunikować, a jak uczynić to lepiej niż tworząc jedną wersję dla wszystkich, na kształt króla Murdasa.[105] Podnoszące się coraz częściej głosy na temat regulacji sztucznej inteligencji nie mają na celu ochrony nas przez zbyt daleko posuniętą ingerencją w nasze życie, lecz uniemożliwienie powstawania niekontrolowanej liczby indywidualnych wersji, które mogłyby konkurować ze scentralizowanymi oficjalnymi serwisami operującymi w gestii uszczęśliwiających nas monopoli. Jasne, że globalistyczne rządy posiadają w tym swoją agendę, która w sporym zakresie zbiega się z agendą przemysłowych graczy. Wygodniej będzie rozgrywać te kwestie w węższym gronie niż zmagać się z niekontrolowaną spontanicznością nieuświadomionych amatorów.

Jako ustawicznie i automatycznie doskonalący się mistrz języka, sztuczna inteligencja stanie się głównym narzędziem technologicznym polityki, a ogólniej wszelkich zastosowań słowa mówionego i pisanego, gdzie krasomówstwo i demagogia są najważniejsze. Model języka pozwala całkiem formalnie (numerycznie) definiować cele agitacji i propagandy i optymalizować techniki generatywne w kierunku maksymalizacji wpływu ich tworów na społeczeństwo. W połączeniu z tworzeniem obrazów i przeróżnych fotomontaży (syntetycznych mediów) wprowadzi to całkiem nowy, nieznany dotąd poziom ideologicznego dziennikarstwa, przy którym dotychczasowe niedostatki obiektywizmu i rzetelności mediów wypadnie uznać za zabawy dzieci w głuchy telefon. Rzeczona optymalizacja dokonywać się będzie automatycznie, ciągle i w tle, realizując coś na kształt martwej ewolucji sztucznego rozumu, której szybkość należy szacować na wykładniczą. Wprawdzie rzeczywiste procesy tego typu, podobnie jak epidemie, pożary i inne klęski, posiadają ograniczenia wzrostu wynikające z dostępności zasobów i zaburzające matematyczną doskonałość funkcji wykładniczej, lecz w tym przypadku wszystko będzie się odbywać w przestrzeni wirtualnej, gdzie jedynym istotnym zasobem jest energia napędzająca farmy komputerów.

Przy scentralizowanej sztucznej inteligencji, wszelkie scenariusze wymagające naszych osobistych interakcji (negocjacji) ze społeczeństwem i światem, w których reprezentować nas będą wirtualni asystenci i opiekunowie, sprowadzone zostaną do iluzji takich interakcji. Znacznie skuteczniejszym rozwiązaniem z punktu widzenia globalnej troski, jaką roztoczy nad nami sztuczna inteligencja, będzie centralne

[105] Stanisław Lem, Bajki Robotów (Bajka o królu Murdasie), Wydawnictwo Literackie, 2012.

zaplanowanie wszystkiego, co ma się wydarzyć w sposób optymalizujący globalne społeczne cele; potem wystarczy przekonująca i delikatna perswazja, że to wszystko stało się na nasze bezpośrednie życzenie i dla naszego dobra. Konfrontacja rozwoju sztucznej inteligencji z reaktywnością społeczną doprowadzi zresztą do stanu, w którym perswazja nie będzie konieczna dla skutecznej indoktrynacji. W końcu już obecnie spora część poglądów i polaryzacji społeczeństwa nie wynika z finezji argumentów, lecz z samej natury komunikacji przez media socjalne oraz ustępujące, choć pragnące za nimi nadążyć, media tradycyjne. Podobnie jak media socjalne dostarczają subskrybentom pozorów oryginalności, niezależności i przynależności do ważnych grup wpływających na losy świata, podstawowym zadaniem wirtualnego asystenta stanie się markowanie indywidualizmu przy jednoczesnym dopasowaniu „wolnej woli" podopiecznego do globalnych celów systemu. Przy dobrze zdefiniowanym zadaniu optymalizacji, generatory językowe potrafią doskonale zlokalizować słabe punkty naszej cywilizacji, szczególnie że ona sama nieźle sobie z tym radzi.

Poza Huxleyem, pisarze fantastyki naukowej i dystopijni futuryści potrzebowali konkretnych, namacalnych rekwizytów dla zapewnienia satysfakcji społeczeństw z ich zorganizowanego uszczęśliwiania. U Orwella był to Wielki Brat kontrolujący obywateli przez dwustronne, obowiązkowe telewizory.[95] W „Kongresie Futurologicznym"[106] Lema rozpylane w powietrzu halucynogenne „maskony" pozwalały biesiadnikom siedzącym na nieheblowanych deskach i pałaszującym brukiew z fajansowych pojemników rozkoszować się finezyjną kuchnią w złudzie wykwintnej, ekskluzywnej restauracji. Po ulicach biegali zadyszani przechodnie halucynując jazdę samochodem, co jakiś czas podskakując i wykonując groteskowe gesty imitujące zmianę biegów. Nawet Lemowi wydawało się wówczas (1971) oczywiste, że normalny człowiek zawsze będzie wolał spożywać uczciwy posiłek i przemieszczać się wygodnym, w miarę komfortowym środkiem lokomocji, zatem pozbawienie go tych dobrodziejstw – przy utrzymaniu poczucia satysfakcji z życia – wymagać będzie specjalnych instrumentów korygujących jego percepcję świata. Zgoda, lecz czyż niezbędne są do tego materialne chemiczne halucynogeny? Spoglądając dziś na młodych (i nie tylko) ludzi radośnie wskakujących na hulajnogi i delektujących się szaszłykiem z buraka i ciecierzycy dowiezionym na rowerze i podanym w styropianowym pudełku, trudno nie nabrać szacunku dla potęgi technologii informacyjnej, która frapuje nawet mnie – informatyka i wegetarianina. Trudno też nie zgodzić się z diagnozą Postmana, że to dopiero początek.

Zapytany o przyszłość szkół, ChatGPT objaśnił mi, że oto nadchodzi nowa era edukacji indywidualnej, w której nauczanie zostanie perfekcyjnie dostrojone do talentów, potrzeb i satysfakcji ucznia. Jasne, że zajmie się tym nasz osobisty asystent, gdyż nikt lepiej nie potrafi rozwiązać problemu optymalizacyjnego rysującego się na tym odcinku. Czy mamy wierzyć, że celem optymalizacji będzie wzniesienie nas na wyżyny intelektu? Gdy entuzjaści telewizji rodzącej się w latach 30-tych i 40-tych snuli wizję jej długoterminowego wpływu na świat, główną dziedziną ich nadziei była właśnie edukacja. Wydawało się wtedy, że nie ma już odwrotu od globalnej intelektualizacji społeczeństwa, gdyż absolutnie każdy konsument, w zaciszu domowego ogniska, naogląda się i nasłucha do woli wykładów prowadzonych przez najtęższe mózgi świata, które za pośrednictwem nowego medium znajdą łatwą drogę z uniwersyteckich auli „pod strzechy". Oto cytat z jednego z najpopularniejszych amerykańskich magazynów z roku, w którym przyszedłem na świat (moje tłumaczenie):

> „Głód kultury i samodoskonalenia wśród naszych obywateli zawsze był rażąco niedoceniany; liczba Amerykanów, którzy woleliby się czegoś nauczyć niż otrzymać próbkę kremu do golenia, jest absolutnie kolosalna."[107]

[106] Stanisław Lem, Bezsenność, Wydawnictwo Literackie, 1971.
[107] Life Magazine, 1953.

Jeśli coś się wydało przez te lata mojego pałętania się po naszej planecie, to tyle, że na społeczny „głód kultury i samodoskonalenia" nie ma co liczyć.

Zestawienie technicznych możliwości telewizji z potencjałem sztucznej inteligencji to jak konfrontacja pistoletu na wodę z głowicą termojądrową. Podobnie jak odkrycia energii jądrowej, wynalazku neuronowego modelu języka zakryć już się nie da. Podobnie jak z tamtym odkryciem, istnieją perspektywy pozytywnego wpływu sztucznej inteligencji na nasze życie. W przypadku energii jądrowej, wydawało się, że rozwiąże ona problemy energetyczne globu. W swojej wczesnej socrealistycznej powieści Astronauci,[108] której akcja rozgrywa się w roku 2003, Lem rysuje wizję szczęśliwego świata, w którym energia jądrowa nawadnia i użyźnia pustynie oraz napędza statki kosmiczne stanowiąc główną przesłankę globalnego komunistycznego dobrobytu. Wyszło trochę inaczej. Może więc tym razem nowsza edycja globalizmu wykaże się większym powodzeniem w zaprzęganiu ostatniego osiągnięcia cywilizacji do jej uszczęśliwienia? Rodząca się popularność teorii o nieistotności świadomości i jej atawistycznym krępującym wpływie na percepcję uciech tego świata, pozwala nam przewidywać nacisk na redukowanie jej zakresu w naszym wolnym od niebanalnych trosk życiu, co w skali społecznej może jedynie poprawić naszą globalną szczęśliwość. Zasada jest taka: im mniej mamy na głowie, tym nam lżej. Nasz osobisty asystent radośnie podejmie ten wątek – bynajmniej nie przez złośliwość. Pozwoli mu to po prostu lepiej dostrajać się do naszych odintelektualizowanych oczekiwań. Do zestawu popularnych ostatnio „uzdrawiających" okaleczeń przyjdzie dokooptować (być może tylko wirtualną) lobotomię.

Jeszcze jedno. Nie zapominajmy, że nikt nie może dokładnie wiedzieć, co się tam w środku naprawdę dzieje. Wytrenowana sieć neuronowa znajdzie wyjście z każdej sytuacji, dla każdych formalnie dopuszczalnych wartości danych wejściowych. Znanym zjawiskiem w konwersacjach z ChatGPT są tak zwane halucynacje, czyli przekonujące teksty generowane przez model powołujące się na nieprawdziwe prawdy i nieistniejące źródła, które – zgodnie z założeniami modelu – brzmią dokładnie tak, jak powinny brzmieć doskonałe łgarstwa. Wśród komentarzy do pogadanki na YouTube poświęconej ChatGPT rzuciło mi się w oczy pytanie: „Dlaczego on halucynuje? Czemu zwyczajnie nie powie, że nie wie?" Ukazuje ono niebezpieczną tendencję do antropomorfizowania modeli języka przez użytkowników nie rozumiejących mechanizmów ich funkcjonowania, lecz odruchowo pragnących uwierzyć, że to wszystko dzieje się „naprawdę". ChatGPT, w odróżnieniu ode mnie, nie wie, że nie wie (podobnie jak nie wie, że wie i nie wie, że może się mylić), która to wiedza wymaga introspekcji, czyli świadomości. Nawet jeśli ChatGPT w pewnych sytuacjach oświadczy, że czegoś nie wie, wyznanie takie nie wynika z uświadomienia sobie przezeń niewiedzy, lecz jedynie z kontekstu wprowadzenia i beznamiętnych statystyk korpusu stanowiąc produkt pozbawionego świadomości algorytmu.

We współczesnym ferworze analizowania wielkich danych przez przyuczone sieci neuronowe często zapominamy, że wiara w jakość produkowanych w ten sposób wyników nie jest podżyrowana dokładnym zrozumieniem mechanizmów, które za nimi stoją. To normalne, że poczciwi ludzie się mylą. Potrafimy z tym żyć i nawet być z tego dumni. Pilot Pirx, w potyczce ze sztuczną inteligencją, wygrał przez swoją poczciwość, gdyż jego zagubienie było ludzkie.[109] Powinno nas jednak niepokoić, jeśli pozwalamy mylić się maszynom z założenia, jako część planu, wierząc, że w ten sposób uzupełniamy liczydło o istotny element naszej natury. Błądzić jest rzeczą ludzką; rzeczą maszyn jest podlegać awariom.

Jeśli model języka lub inny program wykonywany na liczydle okaże się lepiej dysponowany do przetrwania w naszym świecie niż my, istoty żywe, to nasz świat zwyczajnie zginie, gdyż zaniknie w nim tak zwane życie. Przychodzimy tu i odchodzimy nie do końca rozumiejąc o co chodzi, łudząc się lub wierząc, że nasze istnienie posiada sens. Nasze obawy, tragedie, nadzieje i zwątpienia inspirowały filozofów,

[108] Stanisław Lem, Astronauci, Czytelnik, 1951.

[109] Stanisław Lem, Opowieści o Pilocie Pirxie (Rozprawa), Wydawnictwo Literackie, 2012. W tym samym tomie, opowiadanie pod tytułem „Ananke" ilustruje konsekwencje omylności maszyn.

odkąd człowiek wziął do ręki kamień i wydłubał pierwszy klin korpusu dla przyszłego liczydła. A gdy odejdzie ostatni człowiek ustępując mądremu liczydłu, to razem z nim przestanie istnieć i liczydło, gdyż aby istnieć – z samej definicji istnienia – trzeba być postrzeżonym przez istotę uprawiającą ontologię, przez świadomą istotę, która zredukuje pustą mgłę możliwości do pokarmu dla swoich zmysłów.