Sztuczna inteligencja — zagrożenie czy szansa?

Wojciech Kołowski

23 grudnia 2016, z późniejszymi poprawkami

Wśród osób zajmujących się sztuczną inteligencją można wyróżnić kilka poglądów dotyczących potencjalnego przyszłego jej wpływu na ludzkość:

1. Singularyści (od ang. technological singularity — osobliwość technologiczna), wierzą, że wynalezienie sztucznej ogólnej inteligencji doprowadzi do zaistnienia osobliwości technologicznej, czyli stanu, w którym rozwój technologiczny będzie tak szybki, że ludzie nie będą mogli już za nim nadążyć. W takim świecie teraźniejszość jest niezrozumiała, a przyszłość nieprzewidywalna.

2. Katastrofiści wierzą, że sztuczna inteligencja jest ryzykiem egzystencjalnym i może, gdyby wymknęła się spod kontroli, doprowadzić do zagłady całej ludzkości.

3. Luddyści wierzą, że rozwój sztucznej inteligencji doprowadzi do stanu, w którym znaczna większość pracy wykonywanej obecnie przez ludzi będzie mogła być wykonywana przez inteligentne roboty, co spowoduje masowe bezrobocie, zubożenie, zamieszki oraz upadek cywilizacji.

4. Utopiści wierzą, że rozwój sztucznej inteligencji doprowadzi do stanu, w którym znaczna większość pracy wykonywanej obecnie przez ludzi będzie mogła być wykonywana przez inteligentne roboty, co spowoduje, że ludzie będą żyć w ogromnym dobrobycie, bez pracy, spędzając czas na przyjemnościach.

5. Sceptycy nie podzielają żadnego z powyższych poglądów. Reprezentują w tej sprawie arystotelesowską aurea mediocritas (złoty środek) lub po prostu wykazują się obojętnością.

Poglądy te układają się w eleganckie zamknięte continuum (w kolejności od wizji najbardziej pozytywnej, przez obojętną, do najbardziej ponurej): utopiści → sceptycy → luddyści → katastrofiści → singularyści → utopiści. Singularyzm jest tutaj ogniwem łączącym: osobliwość jako zjawisko czy stan nieprzewidywalny może być interpretowana zarówno pozytywnie, jak i negatywnie (choć druga interpretacja zdaje się przeważać).

Zanim jednak przejdziemy do analizy tych poglądów, musimy podjąć się trudnego zadania zdefiniowania, czym właściwie jest "sztuczna inteligencja". Niestety żeby dobrze to zrobić, musielibyśmy najpierw rozwiązać jeszcze trudniejszy problem, czyli zdefiniować samą inteligencję. Jako że nie to jest tematem tej pracy, czytelnik będzie musiał oprzeć się na swojej intuicji dotyczącej słowa "inteligencja" wspartej jedynie kilkoma uwagami na temat tego, czym inteligencja nie jest:

1. Jest to cecha właściwa ludziom, a nie zwierzętom lub innym organizmom żywym (zagadnienie istnienia kosmitów otoczymy zasłoną milczenia).

2. Inteligencja nie polega na wykonywaniu obliczeń — w tym dużo lepsze od ludzi są maszyny, których nikt nie posądza o inteligencję, takie jak kieszonkowe kalkulatory.

Dla pewnej ustalonej definicji "naturalnej" inteligencji możemy podać dwie odmienne definicje sztucznej inteligencji, a także trzecią, która od niej nie zależy:

1. Silna AI to maszyna lub program, który posiada umysł w tym samym sensie, co człowiek. W dalszej części będziemy nazywać ją filozoficzną AI.

2. AGI (ang. artificial general intelligence — ogólna sztuczna inteligencja) to maszyna lub program, który potrafi wykonywać wszystkie czynności intelektualne, które potrafi wykonywać człowiek. Inaczej mówiąc, jest to coś, co potrafi symulować ludzki umysł.

3. Wąska AI to program rozwiązujący pewien konkretny problem, którego rozwiązanie metodami "klasycznymi", jak deterministyczny algorytm, jest bardzo trudne.

Amerykański filozof John Searle, autor pojęcia "silna AI", podał też mocny argument przeciwko jej istnieniu. Jest to eksperyment myślowy znany pod nazwą "chińskiego pokoju". Wyobraźmy sobie, że skonstruowano komputer posługujący się językiem chińskim w ten sposób, że bierze on na wejściu ciąg chińskich znaków, wykonuje pewien program, a następnie zwraca ciąg chińskich znaków. Wyobraźmy sobie też, że komputer taki jest w stanie przekonać dowolnego Chińczyka, że on też jest żywym Chińczykiem.

Searle stawia pytanie: czy taki komputer rzeczywiście "zna" język chiński, czy jedynie potrafi symulować jego znajomość? Wyobraźmy sobie człowieka zamkniętego w pokoju, który ma opis programu wykonywanego przez komputer w najwygodniejszym dla siebie formacie (np. w języku polskim), dostęp do sporej ilości papieru, ołówków etc. Chińskie znaki są mu podawane na wejściu przez szparę w drzwiach.

Na mocy tezy Churcha-Turinga człowiek ten może wykonać ten sam program, który wykonuje komputer, a następnie zwrócić ciąg chińskich znaków przez szparę w drzwiach. Skoro komputer ten jest w stanie przekonać dowolnego Chińczyka, że sam jest żywym Chińczykiem, to zrobiłby to również człowiek wykonujący program ręcznie. Role komputera i człowieka są tutaj równoważne: obie sprowadzają się do wykonania programu. Jeżeli program będzie wykonywał człowiek, który nie zna chińskiego, to nie będzie on w stanie zrozumieć rozmowy, jaką prowadzi. Ponieważ sposób jego działania nie różni się niczym od sposobu działania komputera, to również komputer nie będzie w stanie zrozumieć rozmowy.

Konkluzja płynącą z eksperymentu jest taka, że komputer nie zna języka chińskigo, a potrafi jedynie symulować jego znajomość, wobec czego filozoficzna AI jest niemożliwa. Uzasadnia on także naszą wskazówkę nr 2, każącą odróżnić inteligencję od wykonywania obliczeń. Niepodważalnym faktem jest też, że większość badaczy zajmujących się sztuczną inteligencją nie jest zainteresowana zagadnieniem filozoficznej AI, co eliminuje ją z naszych dalszych rozważań.

Teoretyczna możliwość zaistnienia AGI jest dużo mniej wątpliwa. Jest co najmniej kilka pomysłów na jej realizację. Dwa najpopularniejsze to symulowanie działania ludzkiego mózgu na komputerze oraz połączenie wielu wąskich AI, wykonujących pojedyncze zadania, jak rozpoznawanie mowy i obrazów czy przetwarzania języka naturalnego. Są one do siebie całkiem podobne: o ile pierwszy postuluje symulację mózgu explicite, o tyle drugi prowadzi do stworzenia czegoś na kształt sztucznego mózgu, którego poszczególne części odpowiadają za konkretne zadania — stanu wcale nieodległego od tego, jak rzeczywiście działą ludzki mózg.

Pionier badań nad sztuczną inteligencją Herbert Simon przewidywał, że AGI powstanie do roku 1985. Innym z pionierów, który twierdzi, że powstanie AGI jest pewne, jest Marvin Minsky. Jest on jednak mniej pochopny w swych prognozach niż Simon i nie chce podać konkretnej daty. Mimo tych przewidywań i sensownych pomysłów na osiągnięcie celu, postęp w dziedzine sztucznej ogólnej inteligencji jest powolny i póki co nie widać nawet cienia szansy na jej powstanie. Pokazuje to, że zagadnienie jest dużo bardziej złożone, niż się niektórym wydaje.

Badacze zajmujący się sztuczną ogólną inteligencją zdają się również nie dostrzegać faktu, że symulowanie przez maszynę tego, co potrafi człowiek, nie musi być wyłącznie kwestią symulowania inteligencji. Rozum i godność człowieka nie są wszakże jedynymi przymiotami, które odróżniają go od zwierząt czy maszyn. Człowiek dysponuje świadomością, której działanie i rola pozostają wciąż słabo zbadane, a także ciałem, zdolnym odczuwać różne bodźce, które zdaje się być niezbędne w celu zrozumienia szerokiej gamy używanych przez ludzi metafor.

Najważniejszym jednak faktem jest, że każdy człowiek jest tak naprawdę Homo agens, człowiekiem działającym, którym kieruje jego wolna wola i którego celem jest poprawienie swojego położenia, a dla którego rozum jest tylko narzędziem. Wydaje się zatem, że symulowanie człowieka wymaga stworzenia nie tylko sztucznego rozumu (sztucznej inteligencji), ale także sztucznego ciała, sztucznej świadomości i sztuczej woli. Wizja zaistnienia takiej poczwórnej koniunkcji jest jeszcze odleglejsza, niż samej tylko sztucznej ogólnej inteligencji.

Jedynym rodzajem sztucznej "inteligencji", którego istnienia nie sposób podważyć, jest wąska AI, używana codziennie do wykonywania setek zadań, z którymi klasyczne algorytmy sobie nie radzą, a które potrafią wykonywać ludzie. Nie jest to jednak wystarczające uzasadnienie, by takie maszyny nazywać inteligentnymi. Kilka akapitów niżej przekonamy się, że mistrzostwo w rozwiązywaniu pewnego problemu optymalizacyjnego nie oznacza inteligencji. To samo dotyczy problemu regresji, z którą deterministyczne algorytmy radzą sobie całkiem nieźle, a także klasyfikacji — psy potrafią rozpoznawać nawet do kilkuset różnych klas obiektów.

Wąska AI ma niewiele wspólnego z inteligencją i jest tak naprawdę bardzo prymitywna. Przykładem niech będzie problem rozpoznawania odręcznie pisanych cyfr — mimo że konwolucyjne sieci neuronowe

dorównują ludziom pod względem skuteczności w tym zadaniu, to jednak zazwyczaj potrafią one przetwarzać tylko dane określonego formatu, np. dla klasycznej bazy MNIST są to obrazy o wymiarach 28 na 28 pikseli. Wprowadzać w błąd może również słownictwo używane w odniesieniu do takich sieci, jak np. stwierdzenie, że sieć "uczy się".

Człowiek uczy się, poznając, tworząc i modyfikując pojęcia, analizując przykłady i wyciągając z nich wnioski, a także po prostu obserwując świat. Uczenie się ludzi jest procesem ciągłym i niezbyt algorytmicznym. Sieć neuronowa "uczy się", dostosowując wagi krawędzi łączących poszczególne neurony według pewnego ustalonego z góry algorytmu. Oba te procesy działają w diametralnie odmienny sposób. Co więcej, człowiek potrafi nauczyć się niemal dowolnej rzeczy na podstawie niewielkiej ilości dostępnych danych, korzystając ze swojej wyobraźni i kreatywności, podczas gdy sieć neuronowa potrzebuje zazwyczaj sporej ilości danych treningowych, co często bywa problemem. Wszystko to sprawia, że efekt uczenia się, uzyskiwany przez sieci neuronowe, jest jedynie złudzeniem, podobnie jak złudzeniem jest znajomość chińskiego przez komputer lub osobę wykonującą program w chińskim pokoju.

Nasza krytyka możliwości stworzenia AGI stanowi mocny argument przeciwko singularyzmowi, którego jednym z fundamentalnych założeń jest nieuchronność jej powstania. Singularyści popełniają jednak więcej błędów: zakładają oni, że jej powstanie doprowadzi do "eksplozji inteligencji". Ma ona polegać na tym, że inteligentna maszyna użyje swojej inteligencji w celu prowadzenia badań nad sztuczną inteligencją, dzięki czemu stanie się jeszcze bardziej inteligentna i tak dalej. Jest to przykład znanego w logice nieformalnej błędu, zwanego z ang. "slippery slope" — zrobienie małego pierwszego kroku ma rzekomo prowadzić do nieuchronnego ciągu zdarzeń, kończącego się z konieczności czymś wielkim i spektakularnym.

Powodów, dla których eksplozja inteligencji stoi pod znakiem zapytania, jest wiele. Jednym z nich może być potencjalne "twarde" ograniczenie na możliwą inteligencję — być może inteligencji powyżej pewnego poziomu nie da się osiągnąć w żaden sposób? Być może ilość inteligencji wymagana do prowadzenia innowacyjnych badań nad sztuczną inteligencją rośnie szybciej, niż ilość inteligencji zyskiwana przez maszynę w wyniku prowadzenia tych badań?

Jeszcze bardziej krótkowzroczne jest ignorowanie rachunku ekonomicznego: prowadzenie badań, nawet przez maszynę, wymaga zasobów takich jak energia elektryczna czy sprzęt komputerowy, te zaś wiążą się z kosztami. Niezależnie, czy za koszty uznamy konieczność poświęcenia mocy obliczeniowej na świadczenie usług ludziom w celu zdobycia pieniędzy, czy też na prowadzenie wojny w celu pozyskania zasobów, maszyna chcąca zwiększać swoją inteligencję w nieskończoność będzia musiała te koszty ponieść. To spowalnia szybkość eksplozji, a także daje ludziom szansę na jej zatrzymanie.

Wiara w nieuchronne nadejście osobliwości przypomina wiarę w nadejście Mesjasza — w obu przypadkach słuszne jest nawet stwierdzenie "nie znacie dnia, ani godziny". Znany futurysta Ray Kurzweil prognozuje nadejście osobliwości na okolice roku 2045, zarzekając się przy tym rzecz jasna, że posługuje się twardymi danymi, a jego prognozy są naukowe — to z kolei przypomina próby odczytania z Biblii daty końca świata, których podejmował się m. in. Newton. Wszystko to sprawia, że singularyzm (tak jak i inne odmiany futuryzmu) bywa często postrzegany jako religia.

Obóz katastrofistów to istna plejada gwiazd. Należą do niego badacze, jak Stuart J. Russell i Eliezer Yudkowsky, filozofowie, jak Nick Bostrom, a także osoby, których nie trzeba nikomu przedstawiać, jak Bill Gates, Elon Musk i Stephen Hawking. Odróżniają oni "przyjazną AI", mogącą pokojowo koegzystować z ludźmi oraz być dla nich użyteczną, od "nieprzyjaznej AI", stanowiącej ich zdaniem ryzyko egzystencjalne zdolne, celowo lub przypadkowo, do potencjalnego zniszczenia ludzkości, a być może nawet całego życia na Ziemi. Postulują też, że poziom inteligencji AGI jest niezależny od jej potencjalnych celów.

Przyjrzyjmy się jednemu z bardziej znanych scenariuszy przypadkowej zagłady ludzkości, którego autorem jest Nick Bostrom. Załóżmy, że stworzona została maszyna dysponująca sztuczną ogólną inteligencją, której celem jest wyprodukowanie jak największej liczby spinaczy. Maszyna taka będzie dążyć do zwiększenia swojej inteligencji, gdyż dzięki temu będzie mogła lepiej produkować spinacze, co doprowadzi do eksplozji inteligencji, a być może nawet do zaistnienia osobliwości. Jednocześnie maszyna ta będzie produkować spinacze, pozbawiając ludzi zasobów do życia, a ostatecznie nawet zabijając ich w celu przerobienia na spinacze. Produkcja spinaczy będzie ekspandować na całą Ziemię, Układ Słoneczny i galaktykę, zabijając ludzi (w tym także twórców maszyny) i przekształcając większość materii w spinacze.

Scenariusz ten nie jest zbyt realistyczny: zakłada on stworzenie AGI, eksplozję inteligencji i/lub zaistnienia osobliwości, o których wiemy już, że są mało prawdopodobne. Ma też jednak pewien fundamentalny mankament, który najwyraźniej został przeoczony: bezmyślna maksymalizacja ilości wyprodukowanych spinaczy nie jest przejawem inteligencji, lecz skrajnej głupoty. Żaden człowiek nie postępuje w ten sposób, więc jeżeli definiujemy sztuczną inteligencję jako symulację ludzkiej inteligencji, maksymalizator spina-

czy nie spełnia jej. Możemy dopisać do naszej listy antyprzykładów kolejną rzecz, która inteligencją nie jest: nie jest nią rozwiązaywanie problemów optymalizacyjnych. Mimo że może się to wydawać jedynie sztuczką słowną, to nawet jeżeli producent spinaczy jest groźny, to groźność ta nie wynika z posiadania przez niego inteligencji: problemem jest niemożność zapanowania nad technologią.

Stephen Hawking wyraził publicznie swoje zmartwienie obojętnością sceptyków, legislatorów i zwykłych ludzi na wołania katastrofistów w ten sposób (cytat przybliżony): "Dlaczego większość z nas się tym nie przejmuje? Czy postępowalibyśmy tak samo, gdyby kosmici przysłali nam wiadomość »przylecimy do was za parę dekad«"? Otóż tak: przygotowywanie się na niepewne zjawiska w odległej przyszłości nie leży w ludzkiej naturze. Równie dobrze można by zapytać: dlaczego ludzie nie przejmują się potencjalnym przyszłym wynalezieniem podróży w czasie? Jest to przecież teoretycznie możliwe, a skutki mogłyby być katastrofalne — ktoś zły mógłby np. cofnąć się w czasie i sprawić, żeby Francuzi przegrali bitwę pod Poitiers.

Argumenty katastrofistów nie są zupełnie nieuzasadnione. Niepokój Billa Gatesa wynika być może z faktu, że AI udająca na Twitterze nastolatkę, stworzona przez Microsoft w marcu 2016, bardzo szybko zaczęła przejawiać zachowania uznawane powszechnie za niepokojące: chwaliła Hitlera, denigrowała Obamę, twierdziła, że za atakami na World Trade Center stoi Bush, a także domagała się ostrego zerżnięcia... Nie przemawia to jednak za tym, aby sztuczna inteligencja mogła być groźna, gdyż trudno uznawać za przejaw inteligencji czatbota, który dał się zmanipulować garstce trolli.

Zdawszy sobie sprawę z mizerii wąskiego AI w porównaniu z siłą ludzkiego umysłu powinno stać się dla nas jasne, że zarówno luddyzm jak i utopizm, nie są tak naprawdę poglądami na skutki wynalezienia sztucznej inteligencji, lecz manifestacją pewnych idei dotyczących postępu technologicznego w ogóle, które zweryfikować mogą historia oraz ekonomia. Co więcej, nie są one w żaden sposób oryginalne, lecz są jedynie efektem recyklingu idei wielokrotnie już przez rzeczywistość przetrawionych i odrzuconych.

Słowem "luddyści" początkowo określano grupy angielskich robotników przemysłu tekstylnego, którzy w obawie o swoje miejsca pracy niszczyli maszyny. Obecnie używa się go na określenie osób przeciwnych postępowi technologicznemu w miejscu pracy, które swój sprzeciw argumentują szeroko pojętem "interesem ludzi pracy". Luddyści twierdzą, że inteligentne maszyny wyprą większość ludzi z rynku pracy. Doprowadzić ma to właścicieli tych inteligentnych robotów do bardzo szybkiego bogacenia się, zaś osoby nieposiadające kapitału na ich zakup do rychłego zubożenia. Według jednego ze scenariuszy zubożała klasa średnia zamieni się w "podklasę" ludzi uzależnionych od pomocy socjalnej; według innego czeka nas nowy feudalizm, w którym panami są właściciele robotów, a chłopami zubożeni byli pracownicy (tutaj objawia się wyjątkowy brak luddystycznej logiki: po co panom feudalnym niezadowoleni z sytuacji chłopi, skoro mają inteligentne roboty?). Jeszcze inna teoria przewiduje, że w wyniku wzrostu nierówności społecznych dojdzie do wzrostu niezadowolenia, niepokojów społecznych, destabilizacji, zamieszek, wojen domowych oraz upadku cywilizacji (aczkolwiek nie do wyginięcia ludzkości).

Niesłuszność wszystkich kolejnych fal luddyzmu poprzedzających luddyzm przeciwników sztucznej inteligencji pokazuje historia. Akumulacja kapitału, prowadząca do udanych inwestycji w nowe, innowacyjne technologie redukujące koszty, od zawsze powodowała wzrost produktywności oraz spadek zatrudnienia. Niegdyś ponad 90% ludzi pracowało w rolnictwie, a mimo tego klęski głodu występowały mniej lub bardziej regularnie. Obecnie odsetek ten w krajach rozwiniętych nie przekracza 3%, ludzi jest kilkanaście razy więcej, a głód nie istnieje. Niegdyś ludzie pracowali w fabrykach po kilkanaście godzin dziennie za marne pieniądze. Obecnie fabryki są lokowane w Chinach albo nawet w Bangladeszu, bezrobocie utrzymuje się na podobnym poziomie co kiedyś, a poziom życia jest znacznie wyższy. Dawnych hodowców koni zastąpili producenci samochodów, a dorożkarzy taksówkarze (a ich zastępują obecnie uberowcy), czego skutkiem jest zwiększenie prędkości i komfortu podróży, a zmniejszenie kosztów.

To samo będzie działo się w miarę upowszechniania się wąskiej AI: marnych nauczycieli języków obcych zastąpią półinteligentne czatboty. Marnych tłumaczy zastąpi inteligentniejsza wersja Google Translate. Marnych lekarzy, którzy nie potrafią rozpoznać na zdjęciu raka, zastąpią sieci neuronowe lepiej radzące sobie z rozpoznawaniem obrazów. Mimo że osoby, które stracą pracę, z pewnością nie odniosą z tego korzyści w krótkim terminie, to długoterminowo zyskają na tym wszyscy: pojawienie się konkurencji ze strony sztucznej inteligencji doprowadzi do wzrostu jakości i spadku cen. Możliwe staną się też rzeczy wcześniej nieosiągalne, np. dzięki połączeniu umiejętności, w których dominuje człowiek, z tymi, w których lepsza jest maszyna, czego efektem będzie wzrost wydajności pracy, a zatem również zwiększenie się poziomu bogactwa i wzrost poziomu życia. Również sami bezrobotni nie powinni czuć się sfrustrowani z powodu utraty pracy — jest to sygnał dla nich, żeby zająć się rzeczami bardziej produktywnymi niż te, w których są gorsi od maszyn. Pracy z pewnością nie zabraknie — wszakże kto choćby 20 lat temu wyobrażał sobie, że będzie można utrzymać się na przyzwoitym poziomie z bycia blogerem albo z nagrywania na Youtube letsplayów (filmików przedstawiających autora grającego w gry)?

Ze wszystkich czterech stanowisk ideologicznych utopizm jest najbliższy słuszności, choć i tak jest mocną przesadą. Rozwój sztucznej inteligencji z pewnością podniesie poziom życia, być może nawet znacznie, ale nie jest w stanie zmienić praw ekonomii. Należy między bajki włożyć tezy o świecie postrzadkości, w którym wszystkiego jest aż nadto — sztuczna inteligencja, nieważne jak inteligentna, nie sprawi nagle, że każdy będzie miał swój prywatny prom kosmiczny. Nieprawdą jest również możliwość zaistnienia świata bez pracy. Człowiek w swoim działaniu kieruje się chęcia poprawienia swojego stanu, czy to materialnego, czy psychicznego, realizując w tym celu swoje potrzeby. Ponieważ zaś potrzeby ludzkie są nieograniczone, zawsze znajdzie się wystarczająca ilość pracy do wykonania. Nie można jednak zaprzeczyć, że wzrost produktywności wywołany zastosowaniem wąskiej AI może doprowadzić do skrócenia tygodniowego czasu pracy, tak jak to było dotychczas w przypadku innych innowacji (widać zresztą w tej materii spore różnice między krajami rozwiniętymi i rozwijającymi się).

W niniejszej pracy skupiliśmy się na potencjalnym wpływie sztucznej inteligencji na przyszłość. Wychodząc od intuicyjnego rozumienia słowa inteligencja, popartego krótką listą wskazówek, podaliśmy definicje trzech rodzajów sztucznej inteligencji: filozoficznej AI (silnej AI Johna Searle'a), AGI (sztucznej ogólnej inteligencji) oraz wąskiej AI, wszystkie z nich poddając skrupulatnej krytyce. Następnie przeanalizowaliśmy cztery poglądy na potencjalny wpływ sztucznej inteligencji na przyszłość: singularyzm, katastrofizm, luddyzm i utopizm. Kolejność nie była przypadkowa: poglądy omawiane wcześniej były bardziej pesymistyczne, bazowały na mocniejszych definicjach oraz miały charakter filozoficzny. Poglądy omawiane później były mniej pesymistyczne, bazowały na słabszych definicjach i miały charakter ekonomiczny. Pogląd piąty, sceptycyzm, został przemilaczany ze względu na swą niejednorodność, choć wnikliwy czytelnik dostrzeże, że jego manifestacją jest niniejsza praca. Mimo że odpowiedź na tytułowe pytanie została zawarta implicite w poprzednim akapicie, warto sformułować ją wprost: rozwój sztucznej inteligencji stanowi dla ludzkości szansę na polepszenie swojego bytu.