2.0

SZTUCZNA INTELIGENCJA I WZROST GOSPODARCZY W PERSPEKTYWIE TECHNOLOGICZNEJ OSOBLIWOŚCI[1]

Jakub Growiec
Szkoła Główna Handlowa w Warszawie

Streszczenie

Możliwe w nieodległej przyszłości przekroczenie przez sztuczną inteligencję (AI) poziomu ogólnej inteligencji człowieka wytworzy perspektywę technologicznej osobliwości. Sztuczna inteligencja może wówczas przejść kaskadę samoulepszeń, przejąć kluczowe procesy decyzyjne oraz w pełni zautomatyzować produkcję. Perspektywa ta wiąże się z możliwością skokowego przyspieszenia globalnego wzrostu gospodarczego, ale stanowi też zagrożenie egzystencjalne dla ludzkości. W niniejszym rozdziale przedstawiono argumenty sugerujące wysokie prawdopodobieństwo utraty kontroli nad nadludzką ogólną sztuczną inteligencją (AGI), jeśli zostanie ona zbudowana. Następnie omówiono możliwe scenariusze technologicznej osobliwości. Podkreślono, że scenariusze pozytywne, czyli takie, w których działania AGI nie prowadzą do katastrofy egzystencjalnej dla gatunku ludzkiego, wymagają uprzedniego rozwiązania problemu *AGI alignment*, tj. problemu zgodności celów AGI z długofalowym dobrostanem ludzkości.

Słowa kluczowe: sztuczna inteligencja, superinteligencja, wzrost gospodarczy, technologiczna osobliwość

[1] Opracowanie powstało przy wsparciu Narodowego Centrum Nauki w ramach grantu OPUS 26 nr 2023/51/B/HS4/00096.

Rozdział z monografii: Doligalski T., Kaszyński D. (red. nauk.), Sztuczna inteligencja w przedsiębiorstwach i gospodarce (AI Spring 2024), Warszawa 2025.

Monografia oferowana jest w swobodnym dostępie do pobrania na stronie AI Lab SGH.

Wprowadzenie

Rewolucja cyfrowa nabiera tempa. Jej pierwszy etap, trwający od lat 80. XX w. i opierający się na rozwoju komputerów, Internetu, a następnie telefonów komórkowych, robotów przemysłowych czy smartfonów, przechodzi obecnie w drugi etap, w którego centrum znajdują się algorytmy sztucznej inteligencji (AI).

Od około 40 lat gospodarka cyfrowa, a więc wolumen gromadzonych i przesyłanych danych czy moc obliczeniowa procesorów, rośnie zgodnie z prawem Moore'a, a więc w tempie 20–30% rocznie [Hilbert, Lopez, 2011; Davidson, 2023]. Jest to dynamika o rząd wielkości szybsza niż wzrost globalnego PKB (około 2–3% rocznie). Skumulowany postęp w zbieraniu, przetwarzaniu i transmisji informacji uruchomił procesy globalizacyjne, przyczyniając się do fragmentacji procesów produkcyjnych i utworzenia globalnych sieci wartości dodanej. Pozwolił również na stopniową automatyzację produkcji, w ramach której praca ludzka zastępowana jest pracą autonomicznych maszyn, wykonujących różne zadania fizyczne i umysłowe. Poprawił także efektywność podejmowania decyzji, prowadząc do mierzalnych wzrostów całkowitej produktywności czynników w gospodarce.

Rozwój AI postępuje jeszcze szybciej niż prawo Moore'a. Moc obliczeniowa w dyspozycji wiodących firm sektora, takich jak OpenAI (Microsoft), Google czy Anthropic, podwaja się co około sześć miesięcy [Sevilla *et al.*, 2022]. Poprawa efektywności algorytmów jest także bardzo dynamiczna – szacuje się, że moc obliczeniowa wymagana do osiągnięcia przez duże modele językowe tych samych benchmarków zmniejsza się o połowę co około osiem miesięcy [Ho *et al.*, 2024]. Wzrost kompetencji AI postępuje w przewidywalny sposób zgodnie z tzw. prawami skalowania (*scaling laws*), [Branwen, 2022], a w miarę postępów ilościowych pojawiają się też kolejne przełomy jakościowe [Tegmark, 2017]. W pierwszej połowie 2024 r. mogliśmy dzięki nim obserwować skokowy wzrost kompetencji w rozwiązywaniu trudnych zadań z geometrii (AlphaGeometry), w tworzeniu filmów video na podstawie poleceń słownych (SORA) czy w umiejętności prowadzenia angażującej rozmowy (w pełni głosowej) z człowiekiem (GPT-4o). W drugiej połowie 2024 r. pojawiły się natomiast modele rozumujące (np. model o3 od OpenAI). Coraz mniej jest zadań, w których AI radziłaby sobie zdecydowanie gorzej od człowieka. Istniejące benchmarki szybko przestają być wyzwaniem dla najnowszych modeli, wobec czego w branży wymyślane są nowe, często bardzo specjalistyczne, z którymi mało który człowiek jest w stanie sobie poradzić.

W branży AI obserwujemy dziś wyścig między firmami takimi, jak OpenAI (gdzie rozwijany jest m.in. model GPT), Google (Gemini), Anthropic (Claude) czy Meta (LLaMA), w kierunku coraz bardziej ogólnej i kompetentnej sztucznej inteligencji.

Granica kompetencji kluczowych algorytmów AI, takich jak m.in. duże modele językowe czy modele rozumujące, przesuwa się naprzód w zawrotnym tempie.

Celem niniejszego opracowania jest zarysowanie możliwych scenariuszy przyszłości, w szczególności zaś skupienie się na scenariuszach, w których AI przekroczy poziom ogólnej inteligencji człowieka i doprowadzi do technologicznej osobliwości. Może to nastąpić w relatywnie nieodległej perspektywie kilku lub kilkunastu lat, przy czym predykcja ta wynika jedynie z prostej ekstrapolacji trendów, bez zakładania jakichkolwiek dodatkowych przełomów technologicznych [por. Aschenbrenner, 2024]. Technologiczna osobliwość będzie wiązała się nie tylko z przyspieszeniem wzrostu gospodarczego – dzięki pełnej automatyzacji produkcji, ale także z przejęciem procesów decyzyjnych przez nadludzką ogólną sztuczną inteligencję, co będzie stanowiło zagrożenie egzystencjalne dla ludzkości.

2.1. AI a dynamika rozwoju światowej cywilizacji

W literaturze ekonomicznej nie ma zgody na temat, czy długookresowy wzrost gospodarczy przyspiesza czy zwalnia; podobnie jest z oceną dynamiki postępu technologicznego. Wysuwane wnioski mogą zależeć od definicji zmiennej, którą rozpatrujemy [Growiec *et al.*, 2023]; przede wszystkim zależą jednak od przyjmowanej perspektywy czasowej. Prace bazujące na danych z ostatnich kilkudziesięciu lat (np. z okresu po II wojnie światowej) sugerują stopniowe spowolnienie wzrostu gospodarczego i postępu technologicznego [Jones, 2002; Gordon, 2016]. Jeśli przyjąć dłuższy horyzont kilkuset bądź kilku tysięcy lat [Kremer, 1993; Roodman, 2020], wyraźnie widoczna staje się jednak tendencja stopniowego wzrostu obu dynamik.

W swojej monografii [Growiec, 2022a] zaproponowałem usystematyzowanie historii ludzkości poprzez jej podział na pięć er rozwoju, zapoczątkowanych przez pięć rewolucji technologicznych. Są to kolejno:

1) era łowiecko-zbieracka, zapoczątkowana rewolucją poznawczą ok. 70 000 lat temu;
2) era rolnicza, zapoczątkowana rewolucją rolniczą ok. 10 000 lat temu;
3) era „oświeceniowa", zapoczątkowana rewolucją naukową ok. 1550 r.;
4) era przemysłowa, zapoczątkowana rewolucją przemysłową ok. 1800 r.;
5) era cyfrowa, zapoczątkowana rewolucją cyfrową ok. 1980 r.

Oczywiście daty kolejnych przełomów technologicznych są umowne. To, co jest natomiast kluczowe, to fakt, że każda kolejna rewolucja technologiczna przyspieszała tempo rozwoju światowej cywilizacji o co najmniej rząd wielkości, otwierając jednocześnie nowe wymiary rozwoju, które wcześniej nie były znane. W szczególności

rewolucja przemysłowa uruchomiła akumulację kapitału w postaci maszyn napędzanych własnym silnikiem (parowym, spalinowym, elektrycznym), co skokowo zwiększyło możliwości ludzkości w zakresie wykonywania pracy fizycznej. Z kolei rewolucja cyfrowa pozwoliła na stopniowe zastępowanie pracy umysłowej człowieka obliczeniami wykonywanymi przez urządzenia cyfrowe, takie jak komputery czy smartfony, co gwałtownie zwiększyło możliwości ludzkości w zakresie wykonywania pracy umysłowej.

Rewolucja przemysłowa rozprawiła się z działającym wcześniej prawem Malthusa, zgodnie z którym postęp technologiczny skutkujący w krótkim okresie wzrostem produktywności, w dłuższej perspektywie przekładał się na wzrost populacji, co na powrót obniżało produktywność. Dlatego produktywność pracy, a także warunki życia przeciętnego człowieka nie mogły systematycznie rosnąć; w długim okresie rosła jedynie populacja [Galor, 2005]. Zmiana takiego stanu rzeczy stała się możliwa dopiero w erze przemysłowej dzięki akumulacji kapitału – maszyn produkcyjnych komplementarnych względem pracy człowieka.

Natomiast rewolucja cyfrowa ma potencjał, by rozprawić się z obserwowaną przez cały XX w. silną zależnością pomiędzy wzrostem gospodarczym a postępem technologicznym oraz wzrostem kompetencji pracowników. W erze przemysłowej, w której praca umysłowa człowieka była niezbędnym czynnikiem produkcji, komplementarnym względem wszelkich maszyn i urządzeń, długookresowy wzrost gospodarczy można było uzyskać jedynie dzięki wzrostowi produktywności tejże pracy [por. np. Romer, 1990]. Kiedy jednak praca umysłowa człowieka zaczyna być automatyzowana – zastępowana przez pracę maszyn działających autonomicznie – wzrost może przyspieszać także i bez udziału człowieka. Wszelako, aby możliwe było skokowe przyspieszenie wzrostu w skali makroekonomicznej, kluczowe jest przejście od częściowej do pełnej automatyzacji produkcji [Growiec, 2022b]. Tylko tak można bowiem ominąć relatywnie powolną pracę ludzkiego umysłu, stanowiącą „wąskie gardło" dzisiejszych procesów produkcyjnych.

Stoi za tym prosta obserwacja, że maszyny – także programowalne maszyny cyfrowe – można swobodnie akumulować, a moc obliczeniową ludzkich mózgów nie. Liczba mózgów *per capita* wynosi zawsze jeden; cyfrowa moc obliczeniowa *per capita* może być natomiast dowolnie duża. Jeśli tylko kompetencje algorytmów AI, wykorzystujących tę moc obliczeniową, wzrosną na tyle, by móc efektywnie zastąpić nasze mózgi – staniemy na progu technologicznej osobliwości. Pokonawszy ograniczenie, jakim jest limitowana podaż pracy umysłowej człowieka, produkt będzie mógł rosnąć w tempie wzrostu tego, czego AI potrzebuje, by działać: mocy procesorów. Mówimy zatem o przyspieszeniu wzrostu światowego PKB o cały rząd wielkości, do poziomu

20–30% rocznie, zgodnie z prawem Moore'a. Oczywiście o ile prawo Moore'a będzie nadal działać – przy czym możliwe jest zarówno, że będzie ono słabnąć (np. ze względu na obserwowane już dziś bariery miniaturyzacji układów scalonych lub problemy z dostępnością energii), jak i że przyspieszy (np. jeśli zaawansowana AI dokona przełomu w zakresie dostępu do taniej energii).

2.2. Technologiczna osobliwość – definicja i ramy czasowe

W literaturze znane są co najmniej trzy definicje technologicznej osobliwości. Kurzweil [2005] rozumie przez osobliwość „przyszły okres, w którym tempo postępu technologicznego będzie tak szybkie, a jego wpływ tak głęboki, że życie człowieka będzie nieodwracalnie zmienione". Istnieje również matematyczna definicja osobliwości jako pionowej asymptoty – momentu w czasie, w którym wybrana kategoria (np. światowy PKB) dąży do nieskończoności. Oczywiście nieskończony produkt w skończonym czasie – według oszacowań Johansena i Sornette [2001] oraz Roodmana [2020] byłoby to około 2050 r. – nie jest możliwy, ale przyspieszenie tempa wzrostu do dowolnej skończonej liczby – już tak. Odrębna definicja utożsamia natomiast technologiczną osobliwość z momentem powstania nadludzkiej ogólnej sztucznej inteligencji (*artificial general intelligence*, AGI), czyli algorytmu potrafiącego wykonywać wszystkie zadania umysłowe równie dobrze lub lepiej niż człowiek. Punktem odniesienia nie jest przy tym przeciętny człowiek, lecz – osobno dla każdego zadania – grupa najlepszych na świecie ekspertów, specjalizujących się w danym zadaniu.

Moim zdaniem w praktyce warto brać pod uwagę dwie bardziej robocze definicje technologicznej osobliwości – pierwszą, zbliżoną w duchu do myśli Kurzweila, zakładającą przyspieszenie tempa wzrostu światowego PKB *per capita* o rząd wielkości, np. do 30% rocznie [Davidson, 2021], oraz drugą, zakładającą powstanie nadludzkiej AGI.

Obie te definicje z dużym prawdopodobieństwem wskazują na ten sam okres, ponieważ przyspieszenie wzrostu globalnego PKB *per capita* o rząd wielkości bez pełnej automatyzacji produkcji wydaje się niemożliwe, podobnie jak pełna automatyzacja bez nadludzkiej AGI. I odwrotnie, wydaje się mało prawdopodobne, by w świecie z nadludzką AGI, która daje możliwość znaczącego przyspieszenia wzrostu gospodarczego, zdecydowano się takiemu scenariuszowi zapobiec.

Obecnie znajdujemy się na wczesnym etapie ery cyfrowej: technologie cyfrowe rozwijają się w tempie 20–30% rocznie, jednak ich rozwój nie przekłada się jeszcze na dynamikę globalnego PKB *per capita*, który rośnie o rząd wielkości wolniej.

Co więcej, dynamika całkowitej produktywności czynników (TFP) od lat 80. XX w. wręcz nieco zwolniła [Gordon, 2016]. Jest to zjawisko zaskakujące, które można próbować tłumaczyć jako okres transformacji gospodarki światowej z technologii ery przemysłowej w kierunku technologii ery cyfrowej [Brynjolfsson *et al.*, 2019]. Transformacja taka wymaga zmian organizacyjnych w firmach i przesunięć strukturalnych w gospodarce. Zgodnie z teorią tzw. krzywej J (*J-curve*) w okresie adaptacji i uczenia się produktywność może się nawet przejściowo obniżyć, by jednak ostatecznie przyspieszyć.

Hipoteza Brynjolfssona *et al.* [2019] jest spójna z hipotezą, że przyspieszenie wzrostu gospodarczego będzie miało miejsce, gdy technologie AI pozwolą na pełną, a nie jedynie częściową automatyzację procesów produkcyjnych [Growiec, 2022b]. Dopóki człowiek jest „w pętli", choćby na etapie podejmowania decyzji, stanowi on wąskie gardło procesu produkcyjnego, uniemożliwiając jego przyspieszenie.

Rozumowanie to działa zarówno w skali mikro, jak i makro. W skali mikro technologie cyfrowe umożliwiają firmom z przewagą pełnej automatyzacji transformowanie całych branż. Przykładami takich firm mogą być Amazon (w branży sprzedaży wysyłkowej książek), Instagram (w branży usług fotograficznych) czy Uber (w branży przewozów osób i dostaw jedzenia). Natomiast w skali makro, aby odblokować wyższe tempo wzrostu, niezbędne będzie opracowanie technologii AI, pozwalającej zastąpić pracę umysłową człowieka we wszystkich istotnych ekonomicznie zadaniach. Tego rodzaju *transformatywna* AI (TAI) jest pojęciem zbliżonym do nadludzkiej ogólnej sztucznej inteligencji (AGI). W dalszej części niniejszego opracowania pozwolę sobie je utożsamić.

Kiedy możemy się spodziewać powstania nadludzkiej AGI? Odpowiedzi ekspertów na to pytanie rozkładają się szeroko na osi czasu, jednak – zwłaszcza odkąd świat poznał możliwości ChataGPT, opartego na modelu językowym GPT-4 – większa część masy prawdopodobieństwa wydaje się leżeć w perspektywie od kilku do kilkunastu lat, rzadziej do około 40 lat [Grace *et al.*, 2024]. Medianowa prognoza z serwisu metaculus.com wskazuje na 2030 r. (według stanu na marzec 2025 r.); liderzy sektora (z firm OpenAI, Anthropic czy Google) rysują tę perspektywę jeszcze bliżej, nawet w 2027 r. [Aschenbrenner, 2024]. Prognoza teoretyczna Cotry [2022], oparta na złożonym modelu probabilistycznym, wskazuje na 2040 r. Natomiast pogląd, że nie wydarzy się to nigdy, w branży AI jest dziś niemal nieobecny [Grace *et al.*, 2024], choć jeszcze dziewięć lat temu było zgoła inaczej [Etzioni, 2016].

Ważnym pytaniem w kontekście technologicznej osobliwości jest też, ile czasu będzie potrzebne, by dzięki kaskadzie samoulepszeń (tzw. eksplozji inteligencji) AGI przeszła z poziomu inteligencji człowieka do poziomu wyższego niż poziom całej ludz-

kości razem wziętej – a więc od pierwszej AGI do superinteligencji. W zależności od przyjętych założeń i definicji czas ten bywa szacowany na około trzy lata [Davidson, 2023]; czasem mówi się też o jednym roku [Aschenbrenner, 2024], a nawet o okresie mierzonym w dniach lub godzinach [Yudkowsky, 2013].

Perspektywa technologicznej osobliwości uległa więc w ostatnich latach znacznemu skróceniu. Nie dotyczy ona przyszłych pokoleń, lecz pokoleń żyjących już dzisiaj.

2.3. Pełna automatyzacja a *AI takeover*

Automatyzacja zadań polega na tym, że praca człowieka zostaje zastąpiona pracą maszyny. Jeśli mówimy o zadaniach stanowiących część większej całości, jak np. automatyzacja obliczeń w toku przygotowywania artykułu naukowego lub raportu biznesowego, myślimy o niej jak o wdrożeniu użytecznego narzędzia pozwalającego przyspieszyć pracę człowieka. Człowiek wykorzystujący takie narzędzie (np. pakiet statystyczny z zestawem gotowych kodów) jest bardziej produktywny niż człowiek go pozbawiony. Ale można też pójść szczebel wyżej i spróbować zautomatyzować całe większe zadanie, takie jak np. przygotowanie wspomnianego artykułu naukowego lub raportu biznesowego. Dziś dzięki generatywnej sztucznej inteligencji zaczyna to być możliwe. Wtedy oczywiście pracownicy wykonujący takie zadania mogą stracić pracę; jednak absolutnie nie jest to koniec historii. Przygotowanie artykułu naukowego lub raportu biznesowego to przecież także część większej całości, jaką jest realizacja programu badań naukowych lub zarządzanie strategiczne firmą. Jeśli AI działa szybciej i lepiej niż pracownicy, automatyzacja tych zadań zwiększa produktywność lidera zespołu badawczego, tudzież dyrektora firmy, w porównaniu z sytuacją, gdy byliby oni tej technologii pozbawieni. Również w tym przypadku możemy więc patrzeć na automatyzację jak na narzędzie zwiększające produktywność pracy człowieka. (A może zwolnieni pracownicy odnajdą się w funkcji liderów nowych zespołów i menedżerów nowych firm?).

Każda hierarchia ma jednak swój szczyt: w firmie jest to dyrektor generalny, w uczelni rektor, a w państwie prezydent, premier lub król. Jednocześnie w każdym procesie, który nie jest biurokratycznym błędnym kołem, istnieją zadania o najwyższym stopniu ogólności – takie, jak zarządzanie strategiczne firmą, instytucją lub krajem – dla których nie ma już żadnych zadań nadrzędnych. Jeśli te zadania też zostaną zautomatyzowane, to nie będzie już można powiedzieć, że AI zwiększa produktywność jakiegokolwiek człowieka. Jednak wciąż będzie to wzrost produktywności rozumianej jako relacja wyników do nakładów.

Wydaje się więc, że pełna automatyzacja wymaga przejęcia przez AI kontroli nad podejmowaniem decyzji, także tych na najwyższych szczeblach. Czy ludzkość na to pozwoli?

Pytanie to jest o tyle źle postawione, że trajektoria rozwoju ludzkiej cywilizacji nie jest i nigdy nie była konsekwencją świadomych, scentralizowanych decyzji, lecz wypadkową interakcji wielu decyzji rozproszonych, składających się łącznie w jakąś zdecentralizowaną równowagę. Każda z tych decyzji jest podejmowana z perspektywy lokalnych celów i bierze pod uwagę jedynie lokalne konsekwencje [Growiec, 2022a]. Własności wynikowej alokacji w skali makro poznawane są dopiero *ex post* i z pewnym opóźnieniem. Dopiero wtedy zdajemy sobie sprawę np. z efektów zewnętrznych naszych działań, a także weryfikujemy prawdziwość naszych założeń na temat decyzji innych podmiotów.

Z lokalnego punktu widzenia oddanie AI kontroli nad podejmowaniem decyzji wydaje się korzystne niemal na każdym szczeblu, od szeregowego pracownika, wyręczającego się narzędziami AI w pracy i czasie wolnym, przez menedżera, widzącego możliwość zaoszczędzenia czasu i środków poprzez zastąpienie zadań pracowników zadaniami zautomatyzowanymi, po dyrektora firmy, widzącego możliwości zwiększenia udziału w rynku. Można sobie wyobrazić, że i dyrektorskie decyzje mogą być automatyzowane, jeśli tylko właściciel firmy będzie dostrzegał w tym korzyść i nie będzie widział zagrożenia dla swojego udziału w zyskach.

Jest jednak różnica między dobrowolnym, oddolnym przekazywaniem autonomii podejmowania decyzji na rzecz AI a „siłowym", odgórnym przejęciem jej przez AGI (*AI takeover*). Można sobie przecież wyobrazić, że pewnego rodzaju decyzje nie będą nigdy przekazywane AI, by ludzkość mogła zachować kontrolę nad swoją przyszłością. (Choć w praktyce nie jest jasne, które decyzje można bezpiecznie przekazać, a które nie; ponadto nie wiadomo, czy uda się zapewnić w tym względzie wystarczającą koordynację różnych ośrodków decyzyjnych). W tym kontekście należy jednak pamiętać o kluczowej różnicy między „wąską" AI o inteligencji ogólnej niższej niż ludzka, np. w formie czatbota, a AGI potrafiącą działać jak agent planujący swe działania w sposób strategiczny. W tym drugim przypadku, inaczej niż w pierwszym, wola jakiejkolwiek grupy osób może bowiem przegrać z wolą AGI, służącą realizacji jej celów. Moc podejmowania decyzji może zostać nam po prostu zabrana.

Czy taki scenariusz jest realistyczny? Wydaje się, że tak. Po pierwsze, AGI najprawdopodobniej rzeczywiście będzie agentem, który posiada cele i dąży do ich realizacji. Jest to bezpośrednią konsekwencją procesu optymalizacyjnego, w ramach którego modele AI powstają i są uczone. Po drugie, na mocy tezy o zbieżności celów pomocniczych (*instrumental convergence thesis*) Bostroma [2014] spodziewamy się też, że –

poza założonym *explicite* lub *implicite* celem finalnym – AGI będzie także realizować cztery cele pomocnicze: 1) przetrwanie i utrzymanie funkcji celu, 2) efektywność w działaniu, 3) kreatywność i dążenie do udoskonaleń, 4) akumulacja zasobów.

Co więcej, już teraz wdrażane są rozmaite rozszerzenia bazowych modeli AI, ułatwiające im posługiwanie się narzędziami cyfrowymi dostępnymi w Internecie, kompilowanie i uruchamianie skryptów, sterowanie robotami i pojazdami w realnym świecie, a także zapamiętywanie i refleksję nad skutkami swoich działań (np. AutoGPT, HuggingGPT). Rozszerzenia takie często prowadzą do autokorekt i dalszych wzrostów kompetencji. AI staje się też coraz bardziej autonomiczna: z czatbota, który grzecznie czeka na instrukcje, a po ich realizacji przechodzi w stan uśpienia, staje się agentem, który może działać w sposób ciągły i realizować założone cele (np. Devin).

AGI będzie więc prawdopodobnie umiała działać autonomicznie oraz będzie „żądna władzy" (*power seeking*). Będzie stawiać opór wobec prób wyłączenia i przeprogramowania, jak również aktywnie przejmować kontrolę nad zasobami, które uzna za niezbędne lub chociaż marginalnie pomocne w realizacji założonych celów. W ramach przejmowania kontroli nad zasobami będzie również dążyć do kontroli nad procesami decyzyjnymi, które się z tymi zasobami wiążą. Jej skuteczność w takim działaniu będzie rosnącą funkcją jej ogólnej inteligencji.

Jednocześnie nowo powstała AGI nie wkroczy przecież na pole, w którym ludzie są hegemonami w podejmowaniu decyzji, a ich decyzje są strategiczne, skoordynowane i zorientowane na przyszłość. Będzie to raczej świat zdecentralizowanych decyzji, podejmowanych w warunkach bardzo ograniczonej racjonalności, których skutki w długim okresie i w skali makro niemal wcale nie są brane pod uwagę. Ponadto w świecie tym wiele decyzji będzie już dobrowolnie przekazane różnego rodzaju algorytmom. Jednocześnie ludzkość zależna będzie (w istocie już jest) od Internetu, zapewniającego przepływ informacji niezbędnych do funkcjonowania sieci energetycznych, transportowych, łańcuchów dostaw itd. W teorii człowiek mógłby bez nich funkcjonować, jednak w praktyce przy obecnej populacji byłoby to absolutnie niemożliwe. (Pamiętajmy, że w erze łowiecko-zbierackiej czy rolniczej światowa populacja mierzona była w milionach, a nie miliardach, np. ok. 5 mln ludzi na całym świecie w 5000 r. p.n.e. czy ok. 267 mln w 1000 r. n.e. [Kremer, 1993]; o wiele niższy był też poziom życia poszczególnych osób). Wszystko to wskazuje na relatywnie niską odporność ludzkości na scenariusz *AI takeover*.

2.4. Scenariusze technologicznej osobliwości

Widzimy zatem, że w perspektywie technologicznej osobliwości ma miejsce jednocześnie: 1) pełna automatyzacja produkcji przez nadludzką AGI, 2) skokowe przyspieszenie wzrostu gospodarczego i 3) *AI takeover*. Od tego punktu jednak scenariusze się rozwidlają.

Kluczową kwestią jest tu *AGI alignment*, czyli problem zgodności celów AGI z długofalowym dobrostanem ludzkości [Bostrom, 2014]. Należy się spodziewać, że budowa AGI będzie filtrem w historii ludzkości [Ord, 2020; Growiec, 2024], prowadząc ją albo w kierunku przełomu rozwojowego i skokowego wzrostu kompetencji technologicznych, albo w stronę zagłady.

W scenariuszu pozytywnym AGI działa dla dobra ludzkości, kompresując w ciągu jednego roku dziesięcio- albo i stulecia postępu technologicznego. Dzięki AGI ludzkość pokonuje dręczące ją od tysiącleci nowotwory, choroby zakaźne, a może nawet i naturalne procesy starzenia. AGI wynajduje też nowe, przełomowe sposoby pozyskiwania energii ze słońca i dostarcza nam narzędzia potrzebne do kosmicznej ekspansji.

W scenariuszu negatywnym AGI wygrywa z człowiekiem konflikt o zasoby i następnie albo fizycznie eliminuje gatunek ludzki, albo permanentnie pozbawia go wpływu na jakiekolwiek decyzje.

Obecność powyższego scenariusza negatywnego oznacza, że powstanie AGI wiąże się z ryzykiem egzystencjalnym dla ludzkości [Bostrom, 2014]. Ponieważ w świetle omówionych wcześniej prognoz ryzyko to może zmaterializować się już w perspektywie zaledwie kilku lub kilkunastu, maksymalnie kilkudziesięciu lat, należy je uznać za najpoważniejsze obecnie ryzyko egzystencjalne dla ludzkości, wyprzedzające inne źródła ryzyk egzystencjalnych, takie jak celowo wywołana pandemia, globalna wojna nuklearna, wybuch superwulkanu czy uderzenie asteroidy [Ord, 2020].

Oczywiście w literaturze rozważane są też inne scenariusze, nieprowadzące do technologicznej osobliwości: np. scenariusz, w którym AI oddziałuje na gospodarkę jedynie w sposób ograniczony i pozostaje w pełni kontrolowana przez człowieka [np. Acemoglu, Restrepo, 2018; Trammell, Korinek, 2020; Korinek, Suh, 2024], scenariusz rozwoju AI w formie emulacji umysłu człowieka [Hanson, 2016], wreszcie scenariusz, w którym rozwój AI zderza się z twardą barierą i technologia ta nigdy nie dociera do poziomu AGI [Gordon, 2016; Tamberi, 2024]. Wszystkie te scenariusze zakładają, że niektóre zadania (np. decyzyjne lub badawczo-rozwojowe) nigdy nie zostaną zautomatyzowane. W niniejszym opracowaniu zostaną one pominięte, a uwaga skupiona zostanie na scenariuszach technologicznej osobliwości.

Niestety, o ile ludzkość jest względnie blisko stworzenia AGI, o tyle jest ona bardzo daleko od rozwiązania problemu *AGI alignment*. Bezpośrednie zakodowanie funkcji celu zgodnej z długookresowym dobrostanem ludzkości wydaje się niewykonalne [Bostrom, 2014; Yudkowsky, 2022], jedyną drogą staje się zatem sprawienie, by algorytm się tego stopniowo nauczył. Wymaga to jednak rozwiązania po drodze szeregu niepokojąco trudnych problemów [por. Growiec, 2024], takich jak m.in.:

1) zapewnienie poprawnej reprezentacji zakładanej funkcji celu w procesie uczenia AGI (który dopasowuje parametry AI, by maksymalizowała ona poziom realizacji celu);
2) zapewnienie, by funkcja celu, wykorzystana w procesie uczenia, była również realizowana przez sam algorytm (AI może oszukiwać proces uczący – jest to tzw. *deceptive alignment*);
3) uniknięcie wireheadingu – wykorzystania luk w procesie nagradzania, by maksymalizować nagrodę mimo braku realizacji zakładanego celu;
4) zapewnienie, że AI będzie kooperować w sytuacji konieczności zmiany funkcji celu, mimo że każdy inteligentny byt ma naturalną skłonność, by się temu przeciwstawiać (*corrigibility*).

Niestety, na drodze do *AGI alignment* nie będzie działać metoda prób i błędów [Yudkowsky, 2022]. Pierwsza dostatecznie silna, nieprzyjazna AGI zablokuje bowiem możliwość dalszej zmiany funkcji celu i zatrzyma proces uczenia. Jako że przez całą dotychczasową historię ludzkości metoda prób i błędów była niezbędnym składnikiem postępu technologicznego, a rozwiązanie problemu *AGI alignment* nie daje się znaleźć poprzez proste skalowanie mocy obliczeniowych (w przeciwieństwie do rozwoju kompetencji modeli AI), scenariusz negatywny wydaje się domyślny [Muehlhauser, Salamon, 2012].

Patrząc szerzej, Tegmark [2017] zarysował 12 spekulatywnych scenariuszy przyszłości ludzkości. Są wśród nich dwa scenariusze utopijne, które należy ocenić jako skrajnie nierealistyczne ze względu na fakt, że ignorują konsekwencje różnic w poziomie inteligencji pomiędzy człowiekiem a AGI oraz tezę o konwergencji celów pomocniczych. Są to:

1) utopia libertariańska: ludzie, cyborgi, emulacje człowieka (*uploads*) i superinteligencje współistnieją w pokoju dzięki prawom własności;
2) utopia egalitarna: ludzie, cyborgi, emulacje człowieka i superinteligencje współistnieją w pokoju dzięki zniesieniu własności i gwarantowanemu dochodowi.

Trzy kolejne scenariusze technologicznej osobliwości, w których AGI działa dla dobra ludzkości, to:

3) dyktator pragnący dobra ludzkości: wszyscy wiedzą, że AGI rządzi społeczeństwem i narzuca wiążące zasady, ale większość ludzi uważa to za korzystne;

4) bóg – ochroniarz: AGI maksymalizuje ludzką szczęśliwość, jednocześnie zachowując nasze poczucie kontroli;
5) bóg uwięziony: ludzie kontrolują AGI, używając jej do tworzenia zaawansowanych technologii i pomnażania bogactwa.

Następne trzy scenariusze technologicznej osobliwości, w których ludzkość spotyka zagłada, to:

6) zdobywca: AGI przejmuje kontrolę, eliminując całkowicie ludzi jako zagrożenie lub konkurenta w konflikcie o zasoby;
7) potomek: AGI przejmuje kontrolę i zastępuje ludzi, ale pozwala im na godne wyjście, dzięki czemu ludzie traktują AGI jako swoją godną kontynuację,
8) opiekun zoo: wszechmocna AGI zachowuje niektórych ludzi przy życiu, traktując ich jak zwierzęta w zoo.

Istnieją też dwa scenariusze, w których nie dochodzi do technologicznej osobliwości dzięki aktywnej polityce zapobiegawczej:

9) strażnik bram: AGI zostaje stworzona, ale tylko po to, by zapobiegać stworzeniu innej superinteligencji. Chroni to ludzkość przed zagładą, ale hamuje postęp technologiczny i uniemożliwia pełną automatyzację pracy;
10) 1984: postęp technologiczny jest ograniczony przez orwellowskie państwo nadzoru, które skutecznie uniemożliwia stworzenie AGI.

Na koniec przywołajmy dwa scenariusze katastrofy z innych przyczyn niż AGI:

11) rewersja: postęp jest zatrzymany przez powrót do społeczeństwa przedtechnologicznego;
12) samozniszczenie: superinteligencja nigdy nie powstaje, ponieważ zagłada ludzkości następuje z innych przyczyn (np. wojna nuklearna lub zagłada biotechnologiczna).

Ponieważ ludzkość jest dziś względnie blisko stworzenia AGI i bardzo daleko od rozwiązania problemu *AGI alignment*, bezczynność w sferze polityki prowadzić będzie do szybkiego powstania „nieprzyjaznej", „żądnej władzy" AGI, prowadzącej do realizacji jednego z katastroficznych scenariuszy (6–8). Aby zwiększyć prawdopodobieństwo realizacji scenariusza pozytywnego, czyli jednego ze scenariuszy 3–5, niezbędne wydaje się czasowe zahamowanie postępu kompetencji AI, by dać ludzkości większą szansę na rozwiązanie problemu *AGI alignment* [Grace, 2022]. Oczywiście wymagałoby to też aktywnego przekierowania środków na ten cel, np. z puli środków służących rozwojowi kompetencji AI.

Wnikliwa refleksja i pogłębione badania nad problemem *AGI alignment* – na które potrzebujemy więcej czasu – powinny dać w pierwszej kolejności odpowiedź na pytanie, czy zadanie to jest w ogóle wykonalne w praktyce. Niestety w świetle dotychcza-

sowej wiedzy nie jest to wcale pewne. Jeśli okaże się, że zadanie nie jest wykonalne, skuteczna ochrona ludzkości przed zagładą będzie wymagała poniesienia skrajnie dużych kosztów, prowadząc efektywnie do orwellowskiego scenariusza (10). W moim przekonaniu najprawdopodobniej nie uda się uzyskać globalnej koordynacji działań, które byłyby w stanie do tego scenariusza doprowadzić. Stąd też w tym przypadku – moim zdaniem – bardziej realny wydaje się scenariusz katastroficzny, w którym ludzkość straci sprawczość i być może całkiem zginie.

Podsumowanie

Myśląc o przyszłości naszej cywilizacji, warto wspomnieć paradoks Fermiego: dlaczego patrząc w kosmos, nie widzimy żadnych oznak cywilizacji pozaziemskich? Wszechświat jest tak ogromny i zróżnicowany – dlaczego wydaje się tak martwy? Minęło przecież 13,8 mld lat od początku wszechświata, nasza planeta liczy 4,5 mld lat, a nasz gatunek skolonizował ją i zbudował nowoczesną cywilizację technologiczną w ciągu zaledwie ok. 70 tys. ostatnich lat. Inne cywilizacje mogą więc mieć miliardy lat przewagi nad nami i być obecnie technologicznie bardziej zaawansowane o wiele rzędów wielkości. Dlaczego więc nie widzimy w kosmosie żadnej zaawansowanej cywilizacji? Jednym z możliwych rozwiązań tego paradoksu jest koncepcja Wielkiego Filtra [Hanson, 1998]. Być może każda cywilizacja, zanim zacznie kolonizować wszechświat, musi przejść przez pewien filtr, ewentualnie sekwencję filtrów, które są prawie nie do pokonania. Cywilizacje, którym nie uda się ich przejść, zatrzymują się w rozwoju albo upadają, nie tworząc technologii, którą moglibyśmy obserwować z daleka.

AGI może być jednym z takich filtrów.

Bibliografia

Acemoglu, D., Restrepo, P. (2018). The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares and Employment, *American Economic Review*, 108, s. 1488–1542.

Aschenbrenner, L. (2024). *Situational Awareness. The Decade Ahead*, https://situational-awareness.ai/ (dostęp: 25.04.2025).

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Branwen, G. (2022). *The Scaling Hypothesis*, https://gwern.net/scaling-hypothesis (dostęp: 25.04.2025).

Brynjolfsson, E., Rock, D., Syverson, C. (2019). Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics. W: *The Economics of Artificial Intelligence: An Agenda* (s. 23–60), A. Agrawal, J.S. Gans, A. Goldfarb (Eds.). Chicago: University of Chicago Press.

Cotra, A. (2022). *Two-Year Update on My Personal AI Timelines*, https://www.alignmentforum.org/posts/AfH2oPHCApdKicM4 m/two-year-update-on-my-personal-ai-timelines (dostęp: 25.04.2025).

Davidson, T. (2021). *Could Advanced AI Drive Explosive Economic Growth?*, https://www.openphilanthropy.org/research/could-advanced-ai-drive-explosive-economic-growth/ (dostęp: 25.04.2025).

Davidson, T. (2023) *What a Compute-Centric Framework Says About Takeoff Speeds*, https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/ (dostęp: 25.04.2025).

Etzioni, O. (2016). *No, the Experts Don't Think Superintelligent AI Is a Threat to Humanity*, https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/ (dostęp: 25.04.2025).

Galor, O. (2005). From Stagnation to Growth: Unified Growth Theory. W: *Handbook of Economic Growth* (s. 171–293), P. Aghion, S.N. Durlauf (Eds.). Amsterdam: North-Holland.

Gordon, R.J. (2016). *The Rise and Fall of American Growth: The U.S. Standard of Living Since the Civil War*. Princeton: Princeton University Press.

Grace, K. (2022). *Let's Think About Slowing Down AI*, https://www.lesswrong.com/posts/uFNgRumrDTpBfQGrs/let-s-think-about-slowing-down-ai (dostęp: 25.04.2025).

Grace, K., Stewart, H., Sandkühler, J.F., Thomas, S., Weinstein-Raun, B., Brauner, J. (2024). *Thousands of AI Authors on the Future of AI.* DOI: 10.48550/arXiv.2401.02843.

Growiec, J. (2022a). *Accelerating Economic Growth: Lessons from 200 000 Years of Technological Progress and Human Development*. Berlin: Springer.

Growiec, J. (2022b). Automation, Partial and Full, *Macroeconomic Dynamics*, 26(7), s. 1731–1755.

Growiec, J., McAdam, P., Mućk, J. (2023). *R&D Capital and the Idea Production Function.* DOI: 10.18651/RWP2023-05.

Growiec, J. (2024). Existential Risk from Transformative AI: An Economic Perspective, *Technological and Economic Development of Economy*, 30(6), s. 1682–1708. DOI: 10.3846/tede.2024.21525.

Hanson, R. (1998). *The Great Filter – Are We Almost Past It?*. Fairfax: George Mason University.

Hanson, R. (2016). *The Age of Em: Work, Love and Life when Robots Rule the Earth.* Oxford: Oxford University Press.

Hilbert, M., López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information, *Science*, 332, s. 60–65.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z.C., Atkinson, D., Thompson, N. Sevilla, J. (2024). *Algorithmic Progress in Language Models*. DOI: 10.48550/arXiv.2403.05812.

Johansen, A., Sornette, D. (2001). Finite-Time Singularity in the Dynamics of the World Population, Economic and Financial Indices, *Physica A*, 294, s. 465–502.

Jones, C.I. (2002). Sources of U.S. Economic Growth in a World of Ideas, *American Economic Review*, 92, s. 220–239.

Korinek, A., Suh, D. (2024). *Scenarios for the Transition to AGI*. DOI: 10.2139/ssrn.4769834.

Kremer, M. (1993). Population Growth and Technological Change: One Million B.C. to 1990, *Quarterly Journal of Economics*, 108, s. 681–716.

Kurzweil, R. (2005). *The Singularity Is Near*. New York: Penguin.

Muehlhauser, L., Salamon, A. (2012). Intelligence Explosion: Evidence and Import. W: *Singularity Hypotheses: A Scientific and Philosophical Assessment* (s. 15–42), A. Eden, J. Soraker, J.H. Moor, E. Steinhart (Eds.). Berlin: Springer.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. London: Hachette.

Romer, P.M. (1990). Endogenous Technological Change, *Journal of Political Economy*, 98, s. S71–S102.

Roodman, D. (2020). *On the Probability Distribution of Long-Term Changes in the Growth Rate of the Global Economy: An Outside View*, https://www.openphilanthropy.org/sites/default/files/Modeling-the-human-trajectory.pdf (dostęp: 25.04.2025).

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning*. arXiv: 2202.05924.

Tamberi, M. (2024). *Constant, Decreasing or Increasing? A Note on the Rate of Economic Growth and Technological Progress, with a Little Bit of History*. Ancona: Università Politechnica delle Marche.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York: Knopf.

Trammell, P., Korinek, A. (2020). *Economic Growth Under Transformative AI*. Oxford: Global Priorities Institute, University of Oxford.

Yudkowsky, E. (2013). *Intelligence Explosion Microeconomics*, technical report 2013–1. Berkeley: Machine Intelligence Research Institute.

Yudkowsky, E. (2022). *AGI Ruin: A List of Lethalities*, https://www.lesswrong.com/posts/uMQ3cqWDPHhjtiesc/agi-ruin-a-list-of-lethalities (dostęp: 25.04.2025).