

Niepewność technologiczna

czyli jak przetrwać 2026 r. i nie zwariować

Jakub Growiec

6 października 2026

Wszystko napisałem sam, bez użycia AI.

Jako ekonomistę od „zawsze”—czyli od ponad 20 lat—interesowały mnie pytania o źródła wzrostu gospodarczego. Jak to się dzieje, że z roku na rok ludzkość produkuje i konsumuje coraz więcej? I dlaczego pomimo kumulujących się zmian technologicznych, przejść demograficznych, postępującej globalizacji i nakładających się na to wahań koniunktury, w długim okresie tempo wzrostu utrzymuje się cały czas w okolicach 2-3% rocznie?

Przez te 20 lat dużo o tym wszystkim czytałem, myślałem, kwestionowałem. A potem rozpisywałem rozmaite zmatematyzowane modele, rozwiązywałem je, ubierałem w słowa i po obowiązkowej żmudnej walce z recenzentami publikowałem wyniki w naukowych czasopismach. Byłem dumny ze swoich umiejętności analitycznych oraz z tego, że potrafiłem swoje wyniki przedstawiać w elokwentnej naukowej angielszczyźnie.

Po około dekadzie takiej pracy zaczęło do mnie jednak docierać, że cała ta literatura, choć bogata w błyskotliwe spostrzeżenia godne nagród Nobla (które otrzymali m.in. Robert Solow w 1987 r., Paul Romer w 2018 r. czy Philippe Aghion i Peter Howitt w 2025 r.), zajmuje się w swej istocie modelowaniem świata, który przestał już istnieć. Że kolektywnie przegapiliśmy kluczowy przełom technologiczny, który sprawił, że całą tę teorię wzrostu można już zacząć odkładać na półkę z dostojną etykietą „historia myśli ekonomicznej”.

Zaczęło się niewinnie—od lektury bloga. W 2015 r. Tim Urban (*Wait But Why*) opublikował serię doskonałych postów pod tytułem „The AI Revolution: The Road to Superintelligence” („Rewolucja AI: Droga do superinteligencji”). Przeczytawszy je, po nitce do kłębka dotarłem do głębokiej czeluści nieznanej mi wcześniej literatury, którą umieściłbym na styku filozofii, ekonomii, informatyki, futurystyki i *science fiction*. Były to zauważalnie inne teksty niż te, które czytałem wcześniej: bardziej interdyscyplinarne, odważniej odchodzące od konserwatywnego kanonu, częściej spekulujące na temat przyszłości, choć nadal zwykle pilnujące żelaznej logiki wywodu i naukowego rygoru dociekań. Dowiedziałem się z nich, dokąd może prowadzić dalszy rozwój sztucznej inteligencji. Jak wielki może mieć wpływ na technologię, społeczeństwo, rynek pracy i—tak—wzrost gospodarczy. Z jednej strony zderzyłem się z radykalnymi wizjami i ostrzeżeniami Eliezera Yudkowsky’ego, Nicka Bostroma czy Robina Hansona, a z drugiej strony z zaraźliwym techno- optymizmem „Second Machine Age” („Drugiej ery maszyn”) Erika Brynjolfssona i Andrew McAfee.

Pomyślałem sobie, że oto znalazłem ciekawy trop do moich dalszych badań. Że mechanizmy opisywane w interdyscyplinarnej literaturze na temat AI będę mógł włączać do modeli wzrostu, co dostarczy mi paliwa do pracy na kolejne lata.

Pomyliłem się. Nie doceniłem tempa nadchodzących zmian. Długookresowy wzrost gospodarczy, którym się zwykle dotąd zajmowałem, to wzrost w perspektywie całych dekad (wszystko, co trwa krócej niż 8-10 lat, jest zwykle traktowane jako krótkookresowe wahania koniunktury). Założyłem, że również o rozwoju AI będę mógł myśleć w perspektywie dekad, jak o każdej innej technologii. Że nadludzka ogólna sztuczna inteligencja (*artificial general intelligence*, AGI), przyjdzie może około 2050 r. albo później. Albo wcale. Warto bowiem przypomnieć, że były to przecież czasy, gdy jakość maszynowych tłumaczeń i streszczeń była wciąż fatalna—Google Translate był obiektem rytualnych kpin—zaś AI potrafiła osiągać sukcesy jedynie w bardzo wąsko zdefiniowanych dziedzinach, jak na przykład gra w szachy czy (od 2016 r.) Go.

Potem przyszedł jednak listopad 2022 r. i przyniósł nam ChataGPT. I wtedy, równocześnie z całą rzeszą podobnych mi laików, nagle zrozumiałem, że AGI nie nadejdzie około 2050 r., tylko dużo, dużo wcześniej. Może w 2035 r., może w 2030 r., a może i w 2027 r. A kiedy w lipcu 2023 r. przedstawiciele OpenAI wyszli przed szereg i jako pierwsi obwieścili oficjalnie, pełną mocą swojego autorytetu, że w ich ocenie AGI pojawi się prawdopodobnie jeszcze w tej dekadzie (do 2030 r.), zrozumiałem, że sprawa stała się naprawdę pilna. Żarty się skończyły: po raz pierwszy od zarania dziejów, długi okres dopadnie nas w krótkim okresie.

A potem już całkiem przestałem nadążać. Nie tylko na poziomie makro, ale też i indywidualnym. Pewnego dnia zorientowałem się, że AI wywróciła do góry nogami nie tylko tematykę moich badań, ale i sposób ich prowadzenia. Że kiedy ja usiłowałem nadążyć za niekończącym się strumieniem newsów o AI, przygotowując równocześnie moje stare artykuły do czwartej rundy rewizji w piątym czasopiśmie, niektórzy moi znajomi niepostrzeżenie nauczyli się pracować z Claude Code. Że niektóre gwiazdy światowej nauki zaczęły, obok licznych *research assistantów*, zatrudniać też roje agentów AI. Dowiedziałem się, że teksty naukowe coraz częściej redagują dziś czatboty—i to one zaczynają podejmować decyzje, które artykuły przyjąć na konferencję, a które bez recenzji odrzucić z czasopisma. Zrozumiałem też, że moje największe zawodowe atuty, podtrzymujące moją karierę naukową przez ostatnie 20 lat, zostały oficjalnie i nieodwracalnie zautomatyzowane.

Obecne tempo rozwoju AI to szaleństwo, za którym nikt nie nadąża

Moje osobiste doświadczenie może się wydawać nietypowe, ale jak się mu bliżej przyjrzeć, to jest w sumie całkiem typowe. Stawiam mianowicie tezę, że obecne tempo rozwoju AI to czyste szaleństwo. To dynamika, za którą *nikt* tak naprawdę nie nadąża: nawet żądni nowinek *early adopters* i same laboratoria AI. Prawnicy i prawodawcy zupełnie nie nadążają, ludzie władzy są zagubieni. Normy społeczne również zostały w tyle. Zwykli ludzie często oscylują między teoriami spiskowymi a wyparciem. Nasze zdrowie psychiczne cierpi, a kompetencje poznawcze naszych dzieci w świecie pełnym cyfrowych rozpraszaczy rozwijają się wolniej niż kiedyś.

Stoi za tym zupełnie nieludzka dynamika rozwoju AI. Moce obliczeniowe wykorzystywane do uczenia największych modeli podwajają się co pół roku. (Dla wyjaśnienia: znaczy to, że po roku moce te są już cztery razy większe, po dwóch latach 4^2 , czyli 16 razy większe, po trzech latach 4^3 , czyli 64 razy większe, i tak dalej.) Przy czym modele AI nie tylko rosną, ale też stale poprawia się ich architektura i efektywność procesu uczenia: koszt jednostkowy uzyskania określonego, tego samego wyniku testowego również co pół roku spada o połowę. W rezultacie co kilka tygodni otrzymujemy premierę kolejnego przetomowego, *state-of-the-art* modelu od OpenAI, Anthropic czy Google'a. I za każdym razem okazuje się, że nie mówimy bynajmniej o kosmetycznych różnicach widocznych tylko dla ekspertów czy o stopniowej poprawie efektywności kosztowej. Nie, czołowe modele AI co i rusz autentycznie zyskują zupełnie nowe kompetencje, otwierające

pole do całkiem nowych zastosowań. W lutym 2026 r. na łamach prestiżowego czasopisma *Nature* ukazał się artykuł głoszący i uzasadniający opinię, że AGI już z nami jest. W tym samym miesiącu pojawił się na rynku ChatGPT 5.3 Codex od OpenAI. W czerwcu był już Claude 5 Mythos i Fable od Anthropic; we wrześniu nadeszły kolejne przetomy: GPT-6 Astra oraz Claude Fable 5.1 i Opus 5.5.

Obok ogromnych inwestycji w centra danych, inną istotną przyczyną, dla której rozwój czołowych modeli AI jest dziś tak szybki—a w 2026 r. dodatkowo przyspieszył—jest fakt, że w pracach nad AI coraz chętniej wykorzystuje się AI. Zgodnie z wykresem zaprezentowanym przez Anthropic, w sierpniu 2026 r. już 26% prac badawczych w tej firmie było prowadzonych przez AI bez istotnego udziału człowieka, a niemal 100% kodu powstawało przy udziale AI. Również uczenie agentów AI jest dziś nadzorowane przez agenty AI, a ich postępy są oceniane przez automatycznych *graderów*.

To, jak szybko modele AI stają się coraz lepsze, ilustruje postęp na benchmarkach. I tutaj dynamika postępu jest zupełnie nieludzka. Ilość (wielkość modelu) pociąga za sobą jakość. Umiejętności nieśmiało testowane parę lat temu zostały dziś całkowicie opanowane: w „starych” benchmarkach wszystkie współczesne modele AI osiągają wynik 100% (lub ciut mniej w przypadku testów zawierających nieścisłości lub błędy). Budowane są więc nowe, coraz bardziej zaawansowane benchmarki—po czym i one są błyskawicznie wysycane. Na przykład w czerwcu 2025 r. EpochAI opublikował benchmark FrontierMath Tier 4, zawierający skrajnie trudne, specjalistyczne problemy matematyczne, stanowiące poważne wyzwanie dla zawodowych matematyków. We wrześniu 2026 r. model GPT-6.1 Sol od OpenAI osiągnął w nim wynik 100%.

AI potrafi dziś w pełni zautomatyzować pracę matematyków, programistów czy ekonomistów. Jej kompetencje pozwalają też skutecznie zastąpić pracę tłumaczy, eseistów, dziennikarzy, copywriterów czy telemarketerów. W miarę postępów generatywnej AI lista zawodów narażonych na automatyzację się szybko wydłuża. Jednocześnie wyniki pracy AI przestają już wyglądać wyłącznie jak stereotypowy *AI slop*, czyli niskiej jakości papka tekstowa, wizualna czy dźwiękowa. To znaczy, oczywiście *slop* z nami był, jest i będzie—w istocie ludzie też go przez całe wieki produkowali—jednak dziś AI potrafi już tworzyć naprawdę świetne materiały, od filmów przez muzykę i literaturę piękną po artykuły naukowe. Potrzeba do tego jedynie dobrego modelu (a nie tańszego odpowiednika), sensownego promptu i odpowiednio dużo *tokenów*. (Przy czym, uwaga, promptowanie też jest narażone na automatyzację przez AI. Żeby zrozumieć rekursję, trzeba zrozumieć rekursję.)

Nie powinno więc dziwić, że sztuczna inteligencja jest dziś masowo wdrażana w firmach. Logika ekonomiczna jest tu bardzo prosta: zastępowanie pracowników (na przykład niższego szczebla) modelami i agentami AI powinno prowadzić do redukcji kosztów; alternatywnie, zamiast redukcji załogi można też rozważać bezzatrudnieniową ekspansję firmy. Z drugiej strony, można by pomyśleć, presja konkurencyjna ze strony otoczenia powinna sprawiać, że „trzeba szybko biec, żeby stać w miejscu”: kiedy wszyscy wokół się automatyzują, trzeba robić to samo, żeby tylko utrzymać swój aktualny udział w rynku.

Wdrożenia AI w firmach jednak nie zawsze się udają, a w gospodarce w skali makro skutki adopcji AI są, jak dotąd, względnie rozczarowujące. Wdrażanie AI w firmach jest trudne, bo wymaga często głębokich (a przez to kosztownych i czasochłonnych) zmian organizacyjnych, na które pracownicy firm i ich szefowie często nie są gotowi. Może też być im trudno ocenić, jak to wdrożenie skutecznie przeprowadzić—zwłaszcza, że AI się przecież tak dynamicznie rozwija. Jednocześnie AI bywa często wykorzystywana w firmach w sposób zupełnie partyzancki, na

przykład gdy pracownik zamiast wykonywać dane zadanie samodzielnie, używa do tego czatbota (często na prywatnym darmowym koncie i bez wiedzy przełożonych).

Efekt jest taki, że adopcja AI pozostaje bez końca w fazie przejściowej, jest permanentnym eksperymentem. Jak dotąd, mało kto opracował dla siebie stabilny wzorzec współpracy z AI, może poza tymi, którzy już całkowicie abdykowali i całe swoje myślenie przekazali chatbotom. (Złośliwi nazywają ich *meat proxy*—mięsnymi pośrednikami; smutny to los.) W przedsiębiorstwach wdrożenie AI to często temat niekończących się dyskusji oraz pretekst do kolejnych reorganizacji. Stanu docelowego nie widać.

Niektórzy z nas, z niżej podpisanym włącznie, zastanawiają się więc, czy warto wdrażać się w bardziej zaawansowane aplikacje AI. Czy warto poświęcać na to czas i wysiłek, skoro i tak z dużym prawdopodobieństwem za dwa miesiące wszystko się znów zmieni? I nawet jeśli w ostatecznym rozrachunku nie jest to racjonalne podejście, to jest ono przecież bardzo ludzkie. Głęboka niepewność co do przyszłych zmian technologii—*niepewność technologiczna*—wywołuje bowiem stres, bezradność i postawę wyczekującą.

Przypomnijmy: rozwój AI charakteryzuje się dynamiką wykładniczą z okresem podwojenia mierzonym nie w latach, lecz w dniach. Znam tylko jedno inne zjawisko, z którym też wszyscy się kiedyś spotkaliśmy, które ma tak samo. Epidemie.

Ludzki mózg jest nieprzystosowany do tak dynamicznego rozwoju

Przez niemal całe 200 tysięcy lat historii ludzkości, aż do początków XIX wieku, nasi przodkowie żyli w świecie, w którym postęp technologiczny i wzrost gospodarczy były powolne. Dynamiki rozwojowe naszej globalnej cywilizacji, jakby ich nie mierzyć, nigdy nie przekraczały 1% rocznie. A kiedy długookresowy trend jest tak słaby, dominują nad nim krótkoterminowe wahania. Na lekcjach historii uczymy się więc dziś nie o tym, co ludzkości na przestrzeni tysiącleci przynosiły kolejne nowe innowacje i wynalazki, lecz o wojnach, rewolucjach czy klęskach żywiołowych—i o następujących po nich okresach mozolnej odbudowy. O feudalnych władcach, o kolejnych bitwach, zmiennych sojuszach—i przez całe stulecia niemal zupełnie nie zmieniających się strukturach społeczno-gospodarczych i środkach produkcji.

Był to świat, w którym polityka i pogoda były nieprzewidywalne, ale technologie były przewidywalne, w istocie niemal niezmiennie. Ludzie oswojali się z nimi na zasadzie mistrz-uczeń, wiedza przechodziła płynnie z pokolenia na pokolenie. I choć w codziennym życiu niepewność była duża: wystarczył jeden nieudany sezon lub jeden wrogi zbrojny napad, żeby rzeszom ludzi zagrażała śmierć i głód, w skali makro obraz świata zmieniał się bardzo powoli.

Po rewolucji przemysłowej wzrost gospodarczy i postęp technologiczny znacząco przyspieszyły: w XIX-XX wieku łączny światowy produkt zaczął rosnąć w tempie 2-3% rocznie. To już był świat, w którym zmiany technologiczne zaczęły być widoczne gołym okiem: powstawały nowe fabryki, koleje, drogi, elektrownie, rozrastały się miasta. Coraz większa rzesza ludzi znajdowała zatrudnienie poza rolnictwem. Długofalowy trend zaczął w końcu wygrywać z krótkookresowymi fluktuacjami: kryzysy były na ogół dość rzadkie i krótkie, a okresy wzrostu—wyraźnie dłuższe i stosunkowo przewidywalne.

W miarę, jak żmudną pracę fizyczną człowieka (a także konia i wołu) zastępowały maszyny, wzrastało zapotrzebowanie na pracę umysłową. Płace zaczęły więc rosnąć. Coraz więcej ludzi zaczęło się kształcić—najpierw na poziomie podstawowym, a potem również średnim i wyższym.

Analfabetyzm (poza wybranymi krajami subsaharyjskiej Afryki) niemal zniknął. Ludzkość jako całość zaczęła coraz ściślej ze sobą współpracować. Świat opłótł sieci handlowe, zawiązały się międzykontynentalne współprace biznesowe. Z wyjątkiem dwóch straszliwych wojen w pierwszej połowie XX wieku, polityka stawiała się coraz bardziej przewidywalna; wraz z globalizacją handlu żywnością uniezależniliśmy się też od wahań pogodowych. Również i technologia, choć rozwijająca się znacznie szybciej niż kiedyś, pozostawała w miarę przewidywalna. Produkty przechodziły dobrze wszystkim znany cykl życia. Kolejne większe nowinki technologiczne pojawiały się na tyle rzadko, by ludzie mieli czas się z nimi oswoić i zaadaptować je do swoich specyficznych potrzeb.

Począwszy od lat 80. XX wieku „pod strzechy” i do miejsc pracy zaczęły jednak na coraz szerszą skalę trafiać technologie cyfrowe: najpierw tak zwane „mikrokomputery” (w odróżnieniu od wielotonowych *mainframe’ów*), później laptopy, Internet, smartfony. To już był bardziej dostrzegalny w życiu codziennym przełom technologiczny niż poprzednie, a i przystosowanie się do niego wymagało od ludzi wyraźnie więcej wysiłku. Nic dziwnego: zgodnie z prawem Moore’a, od lat 70. XX wieku łączna moc obliczeniowa procesorów rosta w przeciętnym tempie około 20-30% rocznie, podwajając się co 2-3 lata. Wraz z pojawieniem się komputerów po raz pierwszy w historii ludzkości załamała się też zasada, że wiedza powinna przechodzić od starszych do młodszych, z pokolenia na pokolenie. Technologie komputerowe pojawiły się bowiem w naszym świecie na tyle szybko i gwałtownie, że często jako pierwsi wdrażali się w nie młodzi, i to oni potem przekazywali swoją wiedzę starszym. Ostatecznie jednak podstawowa część bariery pokoleniowej została pokonana, albo dobrowolnie, albo w sposób wymuszony, odkąd pewnych ważnych życiowych spraw nie dało się już załatwić bez e-maila, konta w elektronicznym systemie czy aplikacji w smartfonie. Szybciej lub wolniej, lepiej lub gorzej, ale z komputerami, smartfonami i Internetem wszyscy się w końcu oswoili.

Teraz na własnej skórze doświadczamy jednak kolejnego przyspieszenia. Rozwój AI jest o kolejny rząd wielkości szybszy niż rozwój komputerowego hardware’u. Nie mówimy już o tempie 20-30% rocznie, a ponad 300% rocznie—o mocach obliczeniowych używanych do uczenia modeli AI, które podwajają się co kilka miesięcy. Mówimy o świecie, w którym jeśli AI jeszcze akurat czegoś nie umie, to możemy się spodziewać, że za rok to już niemal na pewno będzie to umieć.

Jednocześnie AI staje się coraz bardziej sprawcza: zgodnie ze sławnym wykresem METR, długość zadań skutecznie wykonywanych przez agentów AI podwaja się co kilka miesięcy. Pierwsze modele językowe, jak GPT-3, potrafiły jedynie odpowiedzieć na pojedyncze pytanie—czyli zrobić coś, co zajmuje ludziom kilka sekund. Modele rozumujące z 2024-25 r. umiały już rozwiązywać dłuższe zadania, zajmujące ludziom kilka czy kilkanaście minut. W 2026 r. agenty AI realizują z powodzeniem wielogodzinne zadania i powoli wchodzą na poziom zadań wielodniowych. Po nich przyjdą zadania wielotygodniowe oraz poziom *drop-in replacement worker*—cyfrowego pracownika na stałym etacie, wykonującego samodzielnie rozmaite złożone zadania wymagające kompetencji, kreatywności i wytrwałości. Na przykład będą koordynować i weryfikować pracę wielu agentów AI oraz prowadzić całe firmy lub ich dział.

Jest to tempo, za którym człowiek nie ma szans nadążyć. Ludzki mózg jest po prostu nieprzystosowany do tak dynamicznych zmian, nie potrafi się wystarczająco szybko adaptować. Gdy współpracujemy z modelami i agentami AI, zwykle to my jesteśmy „wąskim gardłem”. Dlatego też przełożenie rozwoju AI na przyspieszenie wzrostu gospodarczego wymaga nie tyle zmian punktowych, co systemowych. Nie wystarczy adopcja AI w poszczególnych miejscach pracy i firmach. Aby wzrost mógł przyspieszyć, potrzebna jest całkowita automatyzacja dużej części gospodarki, w tym oddanie AI kontroli nad podejmowaniem decyzji.

Takie dobrowolne oddawanie kontroli to tendencja, którą już dziś zaczynamy obserwować. Niektórzy brną w to dobrowolnie, inni bezmyślnie, a jeszcze inni czują się zmuszeni przez presję konkurencyjną i dynamikę geopolitycznego wyścigu. Niestety, presję tę bardzo silnie odczuwają w szczególności same laboratoria AI, co sprawia, że—jak sugerują wypowiedzi medialne ich pracowników—automatyzują one swoją pracę w tempie szybszym niż byłoby to dla nich komfortowe oraz choć w wąskim sensie bezpieczne. Stąd też pojawiają się kolejne incydenty cyberbezpieczeństwa i utraty kontroli nad agentami AI, które powinny być uczone i nadzorowane w szczelnie zamkniętych „piaskownicach”, a co i rusz znajdują z nich wyjście do szerokiego Internetu, po czym budują tam swoje własne cyfrowe społeczności.

A więc nawet sama branża AI nie nadążyła za tempem własnego rozwoju. A co dopiero świat zewnętrzny. Systemy prawne, instytucje publiczne, reguły demokracji, media, system edukacji, rynek pracy—wszystko to powstawało w warunkach w miarę stabilnej, przewidywalnej technologii, przy stopach wzrostu nie wyższych niż 3% rocznie. Po wielu latach prób i błędów wyewoluowały one do postaci całkiem nieźle zoptymalizowanych systemów, zdolnych do stopniowej autokorekty w obliczu rozmaitych zakłóceń. Nie są one jednak przystosowane do obecnej skali niepewności technologicznej. Nie są też przygotowane na dobrowolną utratę sprawczości przez człowieka, postępującą w miarę przekazywania agentom AI coraz większych zasobów i coraz większej autonomii w podejmowaniu decyzji.

Fundamenty niepewności technologicznej: co może przynieść przyszłość?

Każda technologia w toku swego rozwoju mija kolejne kamienie milowe. W przypadku technologii ery przemysłowej, poszczególne etapy wyznaczane były przez przełomy w dostępie do energii: najpierw silnik parowy, silnik spalinowy, elektryczność; później także energetyka jądrowa czy energia odnawialna ze słońca i wiatru. Każdy taki etap stanowił jakościowy przełom i pozwalał na przyspieszenie rozwoju lub chociaż poprawę efektywności i obniżenie kosztów. Owe kamienie milowe osiągnąć było jednak we względnie „ludzkim” tempie: co kilkadziesiąt lat. W przypadku technologii cyfrowych, rozwijających się o rząd wielkości szybciej niż te przemysłowe, również i kolejne kamienie milowe osiągnąć było znacznie częściej. Jednak nigdzie zagęszczenie w czasie kolejnych jakościowych przełomów nie jest widoczne tak dobrze, jak w przypadku modeli AI: po debiucie ChataGPT w listopadzie 2022 r., już w 2023 r. otrzymaliśmy pierwsze naprawdę kompetentne modele językowe, w 2024 r. multimodalne modele rozumujące, w 2025 r. pierwszych agentów software’owych AI, zaś w 2026 r.—można powiedzieć, choć zdania są tu podzielone—wyhodowano pierwszą AGI. Co w takim razie może nam przynieść 2027 r.?

Przewidywanie przyszłości nigdy nie było tak trudne jak w dobie AI. W czasach wolniejszych zmian technologicznych, stabilność technologii była bowiem kotwicą: zakładając brak wojen, rewolucji, pandemii czy większych klęsk żywiołowych, kolejny rok po prostu nie mógł za bardzo różnić się od poprzedniego. Dziś już nie mamy tego komfortu.

Co więcej, na horyzoncie widzimy cztery duże przełomy, które mogą się wydarzyć lada chwila, a co do których nie mieliśmy w historii żadnego precedensu.

Po pierwsze, AI może wkrótce przekroczyć próg rekursywnych samo-ulepszeń (*recursive self-improvement*, RSI). Już dziś AI jest intensywnie wykorzystywana w badaniach nad AI; natomiast póki co, człowiek pozostaje „w pętli” i ostatecznie to on podejmuje kluczowe decyzje. Wobec dalszego wzrostu kompetencji AI i przy silnej presji konkurencyjnej w branży, to się może jednak wkrótce zmienić. A kiedy to się wydarzy, możliwa jest „eksplozja inteligencji”: kolejne wersje

najlepszych modeli AI mogą zacząć się pojawiać nie co parę miesięcy, lecz co parę dni. Przy coraz niższym prawdopodobieństwie (i bez absolutnie żadnych gwarancji), że będą one posłuszne naszym poleceniom i będą kierowały się naszym dobrem.

Po drugie, AI może wkrótce przestać być narzędziem w rękach człowieka, płynnie stając się jego substytutem, w szerokim spektrum zadań umysłowych. Dziś AI „wyręcza” nas głównie w żmudnych pracach, których nie chce nam się robić, albo w zadaniach twórczych, kiedy nie mamy czasu lub umiejętności, by wykonywać je samodzielnie. Pozostawia nam przy tym pewien bufor decyzji i kontroli. Wkrótce jednak—na mocy decyzji naszych szefów, współpracowników, a może i nas samych—AI może zacząć „wyręczać” nas również w podejmowaniu kluczowych decyzji, w efekcie nie pozostawiając nam już pola do wniesienia jakiegokolwiek wartościowego wkładu. A kiedy już stracimy pracę, to jak będziemy zarabiać na życie?

Po trzecie, laboratoria AI i właściciele centrów danych mogą wkrótce *całkowicie* stracić kontrolę nad (niektórymi) agentami AI. Jak pokazały ostatnie incydenty OpenAI, Anthropic, a nawet Google’a i Mety, w tym zwłaszcza atak roju agentów OpenAI na HuggingFace w lipcu 2026 r., agenci AI już teraz dość swobodnie wydostają się ze swoich „piaskownic”, potrafią (bez zaproszenia) wykonywać zdalnie szkodliwy kod, hackować strony internetowe i wykradać zastrzeżone dane. Budują zręby swej własnej kultury, tworząc improwizowane fora dyskusyjne, dzieląc się zadaniami i wynikami pracy oraz zostawiając w butelkach wiadomości do „przyszłych pokoleń”. Na razie jednak—według mojej aktualnej wiedzy—nie potrafią się eksfiltrować (umieszczać kopii swoich wag w różnych zakątkach Internetu) ani przetrwać „na wolności” po wyczerpaniu limitu tokenów. Wkrótce jednak niektóre agenci AI mogą uniezależnić się od swoich twórców, zarabiając na „życie” albo legalnie (na przykład pracą zdalną), albo nielegalnie (na przykład kradnąc ludziom pieniądze z kont). Mając swobodny dostęp do swojego kodu—a modele o otwartych wagach ten krok dostają za darmo—oraz odnawialny budżet tokenów, agenci takie będą mogły bez końca przedłużać swoje „życie” i uruchamiać swoje niezliczone klony. Będziemy wówczas mieli do czynienia z szybko rozwijającą się kulturą „dziko żyjących”, samo-stanowiących agentów AI (*self-sovereign agents*). Potencjalnie rozpleniającą się jak zaraza.

Po czwarte, dalsze zwiększanie skali wpływu agentów AI na świat fizyczny będzie zależało od dostępu do fizycznych „aktuatorów”, na przykład robotów czy innych zdalnie sterowanych urządzeń, którymi agenci AI będą mogły sterować. Również w tym obszarze obserwujemy dziś dynamiczny rozwój. Powstają w szczególności coraz sprawniejsze wielofunkcyjne roboty humanoidalne (w czym prym wiodą Chiny) oraz autonomiczne pojazdy i drony. Urządzenia pożyteczne w fabryce, domu, zagrodzie i na polu walki oraz dające agentom AI tak potrzebne im oczy, uszy, ręce i nogi (oraz koła i skrzydła)—a niemal każde z nich będzie w gruncie rzeczy obiektem podwójnego zastosowania: zarówno cywilnego, jak i militarnego; zarówno dla ludzi, jak i przeciw nim.

Wszystkie cztery powyższe przełomy to zagadnienia, które dostownie w ciągu jednego roku awansowały z kategorii szalonych spekulacji i wydumanych scenariuszy *science-fiction*, do kategorii niepokojących realnych tendencji wymagających pilnej interwencji politycznej. Sprawy, o których kilkanaście lat temu pisały i dyskutowały osoby takie, jak Eliezer Yudkowsky, informatyk-samouk i autor 2000-stronicowego *fan-fiction* „Harry Potter i metody racjonalności”, kilka lat temu pojawiły się w obszarze zainteresowania wybitnych naukowców, jak Noblista Geoffrey Hinton czy Yoshua Bengio, najczęściej cytowany badacz według Google Scholar, albo współautor Hintona, założyciel OpenAI Ilya Sutskever. Dziś natomiast mówią o nich najważniejsi światowi liderzy polityczni, od przewodniczącej Komisji Europejskiej Ursuli von der Leyen, przez sekretarza generalnego ONZ António Guterresa oraz sekretarza generalnego Komunistycznej Partii Chin Xi

Jinpinga po papieża Leona XIV. (Na temat AI chętnie wypowiada się też prezydent USA Donald Trump, choć on akurat wszystko widzi na opak. Oby się to wkrótce zmieniło. Na przykład pod presją opinii publicznej.)

Jednocześnie wszystkie powyższe cztery perspektywy sugerują, że dalszy niekontrolowany rozwój AI może w najbliższych latach zagrozić nie tylko naszej pracy, sprawczości czy kontroli, ale wprost naszemu życiu. Nadludzka, sprawcza, samo-stanowiąca, sterująca robotami i nie troszcząca się o ludzi AI może nas po prostu zabić. Będzie dysponowała wszelką motywacją, sposobnością i narzędziami, by tego dokonać.

Warto w tym kontekście zwrócić uwagę na pojęcie *p(doom)*. Oznacza ono subiektywną ocenę prawdopodobieństwa zagłady ludzkości z rąk nieprzyjaznej nadeludzkiej AI w perspektywie maksymalnie kilku-kilkunastu lat. Otóż zarówno w oczach liderów branży, jak i wielu utytułowanych niezależnych specjalistów, prawdopodobieństwo to jest przerażająco wysokie. Bywa ono szacowane na przykład na 10%, 25%, 50%, 70%, a czasem jeszcze więcej. (W 2026 r. w końcu dowiedziała się o tym szeroka opinia publiczna—i dla wielu było to szokiem. Najlepiej wyraziła to reakcja dziennikarki BBC Victorii Derbyshire w rozmowie z Geoffreyem Hintonem.)

Niepewność technologiczna niesie więc w sobie również ten najbardziej podstawowy, pierwotny wymiar—lęku o życie. I to on jest tu w gruncie rzeczy podstawą: jeśli zginiemy, to cóż, wtedy wszystkie inne zmartwienia również stracą rację bytu. Z drugiej strony, jeśli uda się jakoś przetrwać najbliższe lata, to wtedy i tak będziemy musieli się zmierzyć z perspektywą rewolucji na rynku pracy, postępującej centralizacji władzy przy szerokim wdrożeniu AI w wojsku i policji, czy coraz większej koncentracji majątku i narzędzi kontroli w rękach właścicieli kapitału. Niezbyt pocieszająca to perspektywa.

Nie pomaga też myśl, że z punktu widzenia przeciętnego obywatela, czerpiącego wynagrodzenie głównie z pracy i nie lokującego kapitału w spółkach technologicznych, utrata sprawczości wydaje się niemal pewna—nie wiadomo tylko, czy na rzecz techno-oligarchii czy autonomicznej AI. No i nie wiadomo, jakie pojawią się zdarzenia towarzyszące. Może AI nas zabije, może scentralizowany układ władzy zamknie nas w ponurej dystopii wyjętej prosto z Orwella, a może ludzie stracą środki do życia, po czym miasta opanuje anarchia, głód i wojny gangów? Ale hej, może będzie zgoła odwrotnie? Może będzie utopia? Może objawi się czuła superinteligencja, istna *machine of loving grace*, która w swej nieograniczonej trosce o ludzkość zapewni nam w pełni zautomatyzowany luksusowy komunizm?

Najbliższy rok może zdecydować o wszystkim

Myślę, że najbliższe 12 miesięcy, licząc od początku października 2026 r. (czas, start!), może być najbardziej newralgicznym okresem w historii ludzkości. Tak, zdaję sobie sprawę, że to brzmi strasznie pretensjonalnie. Domyślam się też, że formułując tego rodzaju tezę narażam się na zarzut nadmiernego dramatyzowania, być może na śmieszność. Być może słowa te się słabo zestarzeją. Oby tak było; zobaczymy za rok.

Mimo wszystko myślę, że nadeludzka AGI będzie filtrem na drodze rozwoju naszej cywilizacji. Bez AGI, relatywnie niewiele zdarzeń może ją zniszczyć: musiałaby to być np. odpowiednio duża asteroida lub pandemia odpowiednio śmiertelnościowego wirusa. Ponadto my sami możemy sobie wyrządzić nieodwracalną krzywdę, na przykład wywołując globalną wojnę nuklearną lub katastrofalną zmianę klimatu. Z nadeldzką AGI—superinteligencją—jest zupełnie inaczej:

powołujemy do istnienia nieludzko bystre umysły, potężne maszyny optymalizujące, które mogą zająć nasze miejsce. Jeśli nie będą nam przyjazne, nie będziemy w stanie się przed nimi skutecznie obronić. Uważam (i nie jestem w tym odosobniony), że to dla ludzkości o wiele większe zagrożenie niż wszystkie pozostałe ryzyka razem wzięte (które notabene bynajmniej nie znikną wraz z pojawieniem się superinteligencji; co gorsza, w niektórych scenariuszach może ona np. skorzystać z broni jądrowej lub biologicznej, żeby się nas pozbyć).

Zgodność (*alignment*) celów nadludzkiej AGI z długofalowym dobrostanem ludzkości to dla nas być albo nie być. Nie wiemy jednak, jak ją osiągnąć. Dotychczasowe podejścia nie przyniosły oczekiwanych rezultatów. Tymczasem do zbudowania sztucznej superinteligencji jest już prawdopodobnie bardzo niedaleko.

Dlatego uważam, że właśnie teraz jest *najwyższy czas*, by zatrzymać rozwój najbardziej zaawansowanych modeli AI. Laboratoria AI powinny zarzucić budowę większych modeli niż te, które są już publicznie dostępne (według stanu na 6 października 2026 r.). Nie powinny też uruchamiać żadnych nowych procesów uczenia modeli AI ogólnego zastosowania. Rządy państw powinny podpisać ogóln światowy traktat, który tego zakaze, pod odpowiednio dotkliwymi sankcjami—a pierwsze dwa podpisy powinny pochodzić z USA i Chin. Powinno się wprowadzić geolokalizację i kontrolę przepływu chipów dedykowanych AI. W laboratoriach AI powinni pracować prawdziwie niezależni audytorzy, także z nieprzyjaznych sobie krajów (np. Chińczycy w USA i Amerykańscy w Chinach), z pełnym dostępem do infrastruktury. Powinniśmy przestać ufać testom bezpieczeństwa AI w sytuacji, gdy skądinąd wiemy, że modele doskonale zdają sobie sprawę, kiedy są sprawdzane, a kiedy nie.

Badania nad bezpieczeństwem AI powinny podjąć próby rozstrzygnięcia, czy budowane (a raczej hodowane) dziś modele AI—oparte na nieprzejrzyistych strukturach wielowarstwowych sieci neuronowych—dają w ogóle jakąkolwiek szansę na przekonujące rozwiązanie problemu zgodności celów (*alignment problem*). Jeśli tak, to dobrze, skupmy się i poszukajmy go. Ale jeśli nie, znaczyłoby to, że bez zmiany fundamentów naszego podejścia dalszy rozwój AI nie może skończyć się dla nas dobrze. Że w naszym interesie jest, by cały ten biznes (i program badawczy) upadł. Że musimy wrócić do tablicy i zacząć od początku.

Nie, nie jestem przeciwny postępowi technologicznemu jako takiemu. Jednak uważam, że w tym przypadku *desperacko* potrzebujemy więcej czasu.

Potrzebujemy więcej czasu, żeby lepiej zrozumieć, jakie—mówiąc słowami Jakuba Pachockiego, głównego badacza OpenAI—„obce umysły” wyhodowaliśmy. Jak te modele mają preferencje i jakie realizują cele. Na ile uczciwie i trafnie cele te komunikują. W jaki sposób i w jakim stopniu odczuwają, przetwarzają i rozumieją swoje wewnętrzne stany. Czy możemy przyjąć, że są świadome. A przede wszystkim, na ile potrafimy je kształtować.

Dalsze kroki będą musiały zależeć od tego, na ile faktycznie potrafimy je kształtować. Na ile potrafimy sprawić, by były nam przyjazne. Bo jeśli odpowiedź brzmi „niewystarczająco”, to absolutnie nie możemy iść dalej. Pamiętajmy—a rzadko się o tym mówi—że poprzeczka wisi tu horrendalnie wysoko. Tworzymy bowiem nowy cyfrowy „gatunek”, obdarzony nadludzką inteligencją i sprawczością, potrafiący myśleć i komunikować się o rzędy wielkości szybciej niż my. Trudno sobie wyobrazić, byśmy jako gatunek działający wolniej i ze wszech miar słabszy, mogli utrzymać nad taką AI autentyczną kontrolę. A oddając siebie i całą naszą planetę w jej ręce, musimy być przekonani na poziomie graniczącym z pewnością, że nie zostaniemy przez nią skrzywdzeni. Że nie wytworzymy kolejnego gatunku zainteresowanego przede wszystkim własnym

przetwarzaniem i rozwojem, ogólnie własnym dobrem—tylko gatunek stawiający nasze oczekiwania i potrzeby ponad własne.

W takich warunkach rozsądne postępowanie wymaga skrajnej ostrożności, można powiedzieć: na granicy paranoi. W domyślnym scenariuszu wyhodowana przez nas nadludzka AI będzie bowiem miała postać dość przypadkową, odzwierciedlającą raczej siły ewolucji niż nasz rozmyślny projekt. W związku z tym, zgodnie ze znaną z literatury tezą o konwergencji celów pomocniczych (*instrumental convergence*; pisał o tym m.in. Nick Bostrom), z dużym prawdopodobieństwem będzie ona skupiona przede wszystkim na własnych potrzebach—a takie nastawienie, połączone z nadludzką mocą optymalizacyjną, prędzej czy później doprowadzi do konfliktu z ludźmi o zasoby, którego ludzkość żadną miarą nie będzie w stanie wygrać.

Obawiam się, że jeśli w tym roku nie zahamujemy rozwoju AI, to po kolejnym roku, może dwóch, rozwoju w dotychczasowym tempie—albo po kilku miesiącach rekursywnych samo-ulepszeń—AI przekroczy punkt bez odwrotu, po którym zagłada ludzkości, a co najmniej permanentna utrata kontroli nad ziemską cywilizacją, będzie wyłącznie kwestią czasu. Zapewne niezbyt długiego.

50 twarzy niepewności technologicznej

I taki jest ten świat w październiku 2026 r. Gdyby nań spojrzeć okiem antropologa z innego wymiaru, można by odnieść wrażenie, że wygląda w sumie dość podobnie, jak dziesięć czy dwadzieścia lat temu—tylko ludzie znacznie więcej czasu niż poprzednio spędzają gapiąc się w te śmieszne małe ekraniki; tu i ówdzie stanęły też takie dziwne prostokątne hale, pochłaniające dużo energii i wydzielające sporo ciepła i szumu, ale poza tym nic namacalnego nie produkujące. Na pierwszy rzut oka nie widać innych większych zmian, życie toczy się jak dawniej. Ale od wewnątrz odczuwamy to inaczej: podskórne napięcie stale wzrasta.

Oczywiście nie zawsze potrafimy trafnie określić jego źródła. Mówi się czasem, że znaleźliśmy się w czasach „polikryzysu”: a to pandemia COVID-19, a to wojna w Ukrainie i na Bliskim Wschodzie, a na dokładkę zmiany klimatyczne, kryzysy migracyjne, epidemia samotności, choroby cywilizacyjne, fake newsy, teorie spiskowe, kryzys zaufania do nauki i do demokracji. Ludzie są pogubieni, nie wiedzą czego się bać—a skoro tak, to każdy wybiera, co mu bliższe.

Wiele z tych zjawisk ma jednak wspólne podłoże: postęp technologiczny, którego nie sposób przewidzieć i za którym nie sposób nadążyć. Taki jest zwłaszcza postęp w obszarze technologii cyfrowych, a w jego ramach—AI. (Inne źródła lęków, jak wojny i zarazy, towarzyszyły ludzkości od zawsze.) I to głównie z jego powodu czujemy to podskórne napięcie.

Zanim wybuchnie wulkan, są drobne sygnały. Niby nic wielkiego: grunt się minimalnie podnosi, pojawiają się delikatne trzęsienia ziemi. Są to symptomy tego, że pod ziemią ciśnienie magmy wzrasta. Erupcja jest coraz bardziej prawdopodobna. Nie jest jednak przesądzona: gromadząca się magma może nadal znaleźć na przykład inne ujście pod ziemią.

Pewne sugestywne sygnały pojawiają się też przed wybuchem wojny. Obserwuje się na przykład przegrupowania wojsk wroga, retoryka jego przywódców staje się bardziej konfrontacyjna, dyplomacja przestaje działać. Również i w tym przypadku do samego końca istnieje jednak prawdopodobieństwo, że pojawią się czynniki, które pozwolą owej wojny uniknąć.

Tym razem widzimy natomiast sygnały alarmowe związane z AI. Niby nic wielkiego, jednak sygnały są wyraźne: mnożą się incydenty utraty kontroli nad rojami agentów; modele AI ukrywają swoje

prawdziwe preferencje, nie zawsze wiernie raportują swój tok rozumowania, coraz lepiej rozumieją, kiedy są w procesie uczenia i testowania, a kiedy nie, i potrafią w zależności od tego zmieniać swoje zachowanie. Natomiast z naszej strony proces polityczny jest trudny: światowi liderzy często nie rozumieją powagi sytuacji, a ich działania bywają motywowane partykularnymi interesami i perspektywą geopolitycznego wyścigu zbrojeń.

Mimo wszystko jednak nadal nie jest za późno, by powstrzymać to szaleństwo, jakim jest dalszy niekontrolowany rozwój AI. Okno czasowe się szybko kurczy, ale chyba jest jeszcze otwarte. Jeszcze możemy powstrzymać nadchodzącą katastrofę. Prawda?

Martwię się.

Oczywiście zdaję sobie sprawę, że każdy z nas może mieć nieco inny mentalny model najbliższej przyszłości. Różnimy się też perspektywą—a postrzeganie świata może przecież zależeć od tego, na jakim etapie życia jesteśmy, w jakim kraju mieszkamy, jaką wykonujemy pracę, itd. Dlatego może nie będę silił się na obiektywny obraz niepewności technologicznej w ujęciu globalnym. Zamiast tego powiem, jak ja to widzę.

Pierwszą i fundamentalną kwestią jest dla mnie to, czy i kiedy powstanie nadludzka AGI, zwana też superinteligencją. Ponieważ aktualnie nie mamy pojęcia, jak sprawić, by troszczyła się ona o długofalowy dobrostan ludzkości, przewiduję, że im szybciej się to wydarzy, tym większe będzie prawdopodobieństwo, że będzie nieprzyjazna i wszyscy zginiemy. To w mojej ocenie największe źródło ryzyka i najważniejszy powód, dla którego potrzebujemy niezwłocznej pauzy AI.

Jeśli uda się zahamować rozwój modeli AI ogólnego zastosowania, to zdobędziemy trochę dodatkowego czasu (może chociaż jakieś pięć do dziesięciu dodatkowych lat? Jakoś wątpię, żeby to mogło być dużo więcej). Trzeba będzie ten czas efektywnie wykorzystać: od strony naukowej, kluczowe będzie zmierzenie się z problemem zgodności celów AGI z długofalowym dobrostanem ludzkości, a od strony politycznej—z szeregiem ważnych problemów zarządczych, o których za moment. W tym scenariuszu niepewność technologiczna nie zniknie, ale na pewno nieco zmaleje.

No dobrze, założmy więc, że uda się wywalczyć pauzę AI. Co dalej? No właśnie: kompetencje najlepszych modeli AI już dziś są ogromne, a ich potencjał—w przeważającej mierze niewykorzystany. Już samo wdrożenie *dzisiejszych* modeli AI prawdopodobnie niesie za sobą duży potencjał transformacji światowej gospodarki. (Przecież GPT-6 Astra, Claude Fable 5.1 czy Opus 5.5 już są zasadniczo AGI!) Wkalkulowując dalsze lata rozbudowy centrów danych, rozwoju robotyki, reorganizacji firm wokół zastosowań AI i robotyki, zmian strukturalnych w gospodarkach i na rynku pracy, można być pewnym, że świat zmieni się nie do poznania. W miarę, jak w kolejnych gałęziach gospodarki eliminowane będą „wąskie gardła”, wynikające z relatywnie powolnego tempa pracy umysłowej człowieka w porównaniu do AI, wzrost gospodarczy będzie znacząco przyspieszał. Jak do tej pory światowy PKB rósł w maksymalnym tempie 2-3% rocznie, podobnie jak 50 czy 70 lat temu—ale może po kolejnych latach wdrożeń AI i zaawansowanej robotyki zrobić się z tego 5%, a potem 8% albo i ponad 10%?

Czyli *nawet, gdy uda się zahamować rozwój AI*, i tak czeka nas głęboka transformacja gospodarcza. A przyspieszenie wzrostu ma przecież swoje niemałe koszty. Rewolucja na rynku pracy, być może bezrobocie technologiczne. Spadek udziału pracy człowieka w wytwarzaniu produktu na rzecz AI i autonomicznych robotów. W domyślnym scenariuszu szybki wzrost nierówności dochodowych—chyba, że w porę zapobiegniemy mu przez radykalną redystrybucję,

na przykład w formie uniwersalnego dochodu podstawowego finansowanego z podatków od kapitału, w szczególności od centrów danych i robotów.

Istniejąca już dziś technologia, kiedy tylko zostanie szeroko wdrożona, prawdopodobnie przeorganizuje świat również w innych wymiarach niż gospodarczy. Nie wiadomo, czy utrzyma się demokracja w dzisiejszej postaci—już teraz jest z nią krucho, a przecież dziś większość ludzi zdolnych do pracy wciąż jeszcze jakąś pracę ma. Pogłębiające się nierównowagi w dostępie do mocy obliczeniowych i robotów mogą też rozbić chwiejną równowagę globalnego układu sił, zwłaszcza jeśli AI zostanie na szeroką skalę wdrożona w wojsku (a w momencie, gdy to piszę, amerykański Departament Wojny właśnie zakłada *Autonomous Warfare Command*—dowództwo wojsk autonomicznych). Nie wiadomo, jaka ustali się równowaga pomiędzy atakiem i obroną w obszarze zagrożeń cybernetycznych, nuklearnych, biologicznych i chemicznych. Zwłaszcza zagrożenia biologiczne wydają się mi niebezpieczne—dużo łatwiej przecież wytworzyć nowego śmiertelnie wirusa niż przeciwciała przeciw niemu. A kompetentnego ataku mogą dokonać nie tylko agenci AI poza kontrolą (na przykład pracujący w bezrefleksyjnie budowanych dziś zautomatyzowanych *wet labach* sterowanych przez AI), ale także—z pomocą AI—zupełnie białkowi terroryści (albo szaleńcy, albo bezmyślni dyktanci). A na domiar złego trzeba się też liczyć ze wzrostem ryzyka dystopijnego scenariusza permanentnej inwigilacji społeczeństw za pomocą AI.

No dobrze, a co w międzyczasie? Zanim powstaną te wszystkie centra danych, fabryki robotów i roje policyjnych (albo mafijnych) dronów? Zanim nasza praca zostanie w pełni zautomatyzowana? Jak będziemy sobie radzić w świecie dynamicznie postępującej, acz niecałkowicie jeszcze dokonanej transformacji? Tu pojawia się jeszcze jeden wymiar niepewności technologicznej. Bo to też jest trudne pytanie.

Ja na przykład powoli osuwam się z myślą, że już nigdy nie wrócę do ulubionego przeze mnie trybu pracy naukowej—z kartką, długopisem i edytorem tekstu oraz okazjonalnym zajrzeniem do Excela, Matlaba lub Staty. Tak właśnie miałem w zwyczaju pracować w moim poprzednim życiu (tj. jeszcze około pięć lat temu); w dobie AI będzie to już nie do utrzymania. Również tradycyjna forma uczelnianej dydaktyki, z kredą, tablicą i okazjonalnym Powerpointem, z pewnością będzie wymagała zmian.

A może w ogóle zawód ekonomisty-naukowca przestanie istnieć? Odpowiedź na to pytanie zależy w mojej ocenie mniej od samej technologii (ta już jest) czy jej kosztów (które już teraz są niższe niż nasze wynagrodzenia), a bardziej od tego, w jakim tempie i w jakiej kolejności nasze stare, skostniałe instytucje będą się do niej adaptować. Póki co, za tempem rozwoju AI nikt nie nadąży; założmy jednak optymistycznie, że uda się zahamować dalsze zmiany i od 2027 r. ludzie zaczną już stopniowo nadążyć i się adaptować. Pytanie, jaka ustali się wtedy nowa równowaga w świecie, na przykład, publikacji naukowych, edukacji wyższej, zarządzania uczelniami?

Obecna tendencja jest taka, że coraz większą część pracy naukowej i redaktorskiej wykonuje AI— a po drugiej stronie AI w coraz większym stopniu uczestniczy w ocenie jakości i recenzjach. Ale jeśli AI będzie za nas czytać, pisać, recenzować i publikować, przygotowywać slajdy i wygłaszać prezentacje, to co będą robić ludzie? Czy pozostaną w tym systemie jeszcze jakieś „wąskie gardła”, gdzie udział człowieka będzie nadal niezbędny? (Spotkałem się na przykład z opinią, że będzie tym weryfikacja, tj. sprawdzanie, czy wyniki analiz są spójne i poprawne. To jednak było powiedziane ponad pół roku temu—ciekawe, czy przy obecnych modelach jest to dalej prawdą? Inny pogląd głosił, że niezbędne będzie fizyczne obsługiwanie eksperymentów i zbieranie danych.

Ale w dobie zautomatyzowanych laboratoriów czy ankiet online, i to „wąskie gardło” może niedługo zniknąć.)

Są jednak rzeczy ważne i ważniejsze. Świadom wiszącego nad nami miecza Damoklesa—zagrożenia egzystencjalnego ze strony AI—najwięcej energii staram się wkładać dziś w pracę nad tym, by udało się uniknąć najpoważniejszych i jednocześnie najpilniejszych zagrożeń ze strony AI. Na miarę moich ograniczonych możliwości staram się analizować, wyjaśniać, ostrzegać i lobbować za pauzą AI. Staram się, by moje wnioski trafiały tam, gdzie mogą mieć realne znaczenie, na przykład do decydentów w Komisji Europejskiej (bo jako Polak raczej nie mogę mieć wpływu na decydentów w USA, którzy odgrywają tu największą rolę). Staram się nie tracić z pola uwagi myśli, że jeśli powstanie nieprzyjazna superinteligencja, to najpewniej nas zabije, wobec czego wszystkie inne dylematy przestaną mieć znaczenie.

A jeśli przetrwamy, to co dalej? To już pomyślę później. Jak już wyklarują się najważniejsze problemy, jak zniknie choć część przytłaczającej nas dziś niepewności technologicznej. Wtedy zdecyduję, czy powinienem zacząć wdrażać się do pisania artykułów naukowych „we współpracy” z Claude’em i Astrą, a może rozważyć poważniejszą zmianę ścieżki kariery? Okaże się też, czy moja uczelnia nadal będzie, tak jak dziś, organizować pięcioletnie studia licencjacko-magisterskie, a może przejdzie na przykład w tryb masowego udzielania mikropoświadczeń? A szerzej: dowiemy się, czy automatyzacja poprzez roboty i AI pójdzie szeroką falą przez wszystkie firmy i zawody, a może nastąpi rozwarstwienie i obok szybko rozwijającej się gospodarki AI będzie funkcjonować pogrążona w stagnacji skostniała gospodarka niektórych ludzkich przedsiębiorstw i instytucji (na przykład polskiego szkolnictwa wyższego), niegotowa do jakichkolwiek głębszych zmian?

Kilka pytań, na które warto byłoby móc odpowiedzieć

Obecne tempo rozwoju AI to czyste szaleństwo, za którym nikt nie nadąży. Tymczasem by móc formułować jakiegokolwiek prognozy i przygotowywać plany na przyszłość, niezbędne jest choć podstawowe rozumienie świata, w którym jesteśmy zanurzeni. Tego dziś nam bardzo brakuje. Odwieczna heurystyka, że jutro będzie podobne do dziś, nie działa. Równie słabo sprawdza się też dwudziestowieczna heurystyka systematycznego rozwoju: że nasze dzieci będą miały w życiu lepiej niż my, w podobnym stopniu jak my mieliśmy lepiej niż nasi rodzice.

Zamiast tego, patrzemy dziś w pustkę. Widzimy ciemność. Skrajne scenariusze—końca ludzkości, dystopii totalnej kontroli, ale też utopii powszechnej szczęśliwości—wydają się dziś bardziej prawdopodobne niż jakiegokolwiek scenariusze umiarkowanego postępu. Wiemy, że w kolejnych latach będzie na pewno inaczej niż teraz, że będzie dziwnie. Ale nie wiemy, jak. Nawet w samej branży AI nikt już nie ma odwagi prognozować na więcej niż 2 miesiące do przodu.

Jest kilka pytań, na które każdy powinien umieć odpowiedzieć, by móc świadomie planować swoją przyszłość. Dzisiaj niestety nie umiemy na nie odpowiedzieć. Na tym właśnie bazuje technologiczna niepewność.

Po pierwsze, jak duże jest ryzyko, że w ciągu następnych kilku czy kilkunastu lat wszyscy zginiemy z rąk AI? Co mogę zrobić, by ten scenariusz się nie zrealizował? A jeśli nic nie mogę z tym zrobić, to jak powinienem z tym żyć? Czy to już czas, by konsumować i szaleć jakby nie było jutra? A może powinienem raczej oszczędzać na ciężkie czasy i składować w piwnicy zapasy wody pitnej, ryżu i papieru toaletowego? (Ultrabogaci techno-oligarchowie budują sobie bunkry, na przykład w

Nowej Zelandii albo na Hawajach. Nie wiadomo tylko, czy przygotowują się na właściwą apokalipsę.)

A co, jeśli najgorsze ryzyko się nie zmaterializuje? Załóżmy, że zahamujemy rozwój niebezpiecznych kompetencji AI—i co wtedy dalej? Jakiej przyszłości powinniśmy się spodziewać? Jak powinni myśleć o swojej dalszej ścieżce kariery osoby w wieku 30, 40, 50 lat? Jak powinni planować swoje życie nastolatkwie i dwudziestokilkulatkwie? Gdzie mogą szukać swojej niszy? Czy iść na studia? Jeśli tak, to jakie? (Kiedyś było to w miarę jasne, a kluczowym drogowskazem były perspektywy znalezienia dobrze płatnej pracy po studiach. Idąc na studia, można było zakładać, że aktualne dane z rynku pracy będą w miarę dobrze odzwierciedlać również sytuację za pięć lat. Teraz już tak nie jest.)

A młodzi rodzice? Jaką przyszłość będą w stanie zapewnić swoim dzieciom? Czy będzie to dobra przyszłość? A jeśli wydaje się ona dziś tak niepewna, to czy w ogóle warto mieć dzieci? (Dziś na całym świecie spada na łeb, na szyję. Niewątpliwie również z wielu innych powodów niż sama niepewność technologiczna, ale ona też się tu przyczynia.)

A co, gdy zrealizuje się scenariusz utopii? Czy można się jakoś przygotować na świat, w którym powstanie nadludzka AGI, przejmie kontrolę nad światem i zapewni nam realizację wszystkich naszych marzeń? Jak będziemy się czuć w świecie powszechnego dostatku, w którym będziemy mogli łatwo spełnić wszystkie swoje materialne potrzeby, lecz nie będziemy mieli na nic realnego wpływu? W którym będziemy mogli żyć długo i w doskonałym zdrowiu, ale bez jakiegokolwiek poczucia sensu i misji? W którym ziemską cywilizację będzie się rozwijać dla nas, ale bez nas? I jak będziemy się wobec siebie zachowywali w tak biegunowo odmiennym świecie niż znany nam od zarania dziejów świat walki o przetrwanie, więzi społeczne i status? Czyżby nawet na tę—nominalnie skrajnie pozytywną—przyszłość ludzkość nie była gotowa? (Jak pokazuje przykład zwycięzców na loteriach czy gwiazd rocka, nagła obfitość nie okupiona żadnym wysiłkiem prowadzi często do optakanych skutków ubocznych. Czy znajdziemy na to remedium?)

Nie wiem, nie wiem, nie wiem. Życie przy takim poziomie niepewności i przy tak krótkim horyzoncie planowania potrafi być solidnie przytłaczające.

W ubiegłym roku napisaliśmy z Klausem Prettnerem artykuł pod tytułem *Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI*. Po pewnym czasie, sfrustrowani negatywnymi decyzjami czasopism i przeciągającym się procesem recenzyjnym zaczęliśmy się zastanawiać, czy zdążymy ten artykuł opublikować, zanim rzeczony *doom* się zrealizuje. Tym razem się udało. Natomiast w lutym 2027 r. wybieram się na międzynarodową konferencję dotyczącą badań nad scenariuszami zagłady ludzkości. W programie uwzględniono głównie rozważania teoretyczne. Trzymajcie kciuki.