O wizjach „dobrej przyszłości" ludzkości w świecie ze sztuczną superinteligencją

Jakub Growiec

14.01.2026

Wyobraźmy sobie świat ze sztuczną superinteligencją, przewyższającą kompetencje umysłowe człowieka we wszystkich istotnych aspektach: myślącą szybciej i głębiej, lepiej przewidującą przyszłe zdarzenia, znajdującą lepsze rozwiązania wszystkich trudnych zagadek, tworzącą lepsze plany na przyszłość i efektywniej je realizującą. Bardziej sprawną intelektualnie nie tylko niż każdy człowiek z osobna, ale też w porównaniu do całych firm, korporacji, społeczności i społeczeństw. Taką, która nigdy nie śpi i nie choruje. Oraz dysponującą wystarczającą mocą obliczeniową, by móc te kompetencje na dużą skalę realizować.

Taka AI będzie dysponowała potencjałem, by przejąć kontrolę nad wszystkimi kluczowymi decyzjami determinującymi ścieżkę rozwoju światowej cywilizacji oraz los każdego człowieka z osobna.

Obok *potencjału*, superinteligencja najprawdopodobniej będzie też miała *motywację*, by tę kontrolę przejąć. Nawet jeśli nie będzie do tego wprost dążyć, to i tak będzie mieć motywację instrumentalną: niemal każdy cel łatwiej jest bowiem osiągnąć, kontrolując otoczenie—w szczególności eliminując zagrożenia i gromadząc zasoby.[1]

Oczywiście nie wiemy, jak takie przejęcie kontroli będzie przebiegać. Może będzie jak w *Terminatorze*: gwałtowne, całkowite i bezbrzeżnie krwawe? Ale może będzie stopniowe i początkowo niemal niedostrzegalne? Może, jak w okrutnym eksperymencie z gotującą się żabą, nie będziemy dostrzegali problemu tak długo, dopóki nie będzie już za późno? Może AI początkowo pozostawi nam wolność podejmowania decyzji w obszarach, na których będzie jej mniej zależeć, by potem stopniowo nam zakres tej wolności zawężać? Może zrealizuje się mieszanka obu scenariuszy: utrata kontroli będzie najpierw częściowa i dobrowolna, by znienacka przeistoczyć się w zmianę permanentną i przymusową?

A może—wyobraźmy sobie—będzie to zmiana, która w ostatecznym rozrachunku będzie dla nas *korzystna*?

Spróbujmy sobie odpowiedzieć, jaka mogłaby być „dobra przyszłość" ludzkości w świecie kontrolowanym przez sztuczną superinteligencję. Jakie cele powinna by ona realizować, by móc gatunkowi ludzkiemu tę „dobrą przyszłość" zagwarantować? W jakich warunkach moglibyśmy powziąć przekonanie, że będzie ona działała w naszym imieniu i dla naszego dobra?

Aby odpowiedzieć na te pytania, musimy zrobić krok wstecz i zastanowić się, do czego my sami dążymy—i to nie tylko każdy z nas z osobna, ale też jako cała ludzkość razem wzięta.

1/ Ścieżkę rozwoju cywilizacji wyznaczają zmiany technologiczne

Mistrz Oogway z filmu „Kung Fu Panda" powiedział w swej żółwiej mądrości, że „przeszłość to historia, jutro to tajemnica, ale dzisiaj to dar losu". Niektórzy odczytują to jako sugestię, by po prostu nie martwić się i żyć chwilą. Ale gdy rozłożyć to zdanie na czynniki pierwsze, można je

[1] Zjawisko to nazywamy konwergencją celów pomocniczych. Sporo pisał o nim m.in. Nick Bostrom (2014).

odczytać zgoła inaczej. Kluczem jest tu ciągłość między poszczególnymi okresami. Przeszłość (historia) uruchomiła bowiem procesy, które działają także dziś. Procesów tych—technologicznych, społecznych, gospodarczych czy politycznych—nie możemy odwrócić ani zatrzymać, ale możemy je na bieżąco obserwować oraz do pewnego stopnia kształtować, choć będą one też podlegały niezrozumiałym dla nas, być może losowym zmianom (dar losu). Prawdopodobnie będą one oddziaływały na nas także jutro, jednak nie wiemy, jak (tajemnica). Może więc naszym zadaniem nie jest bezrefleksyjnie żyć chwilą, ale wręcz przeciwnie—próbować zrozumieć wszystkie te długofalowe procesy, by móc je lepiej przewidywać i lepiej nimi sterować? Wykorzystać nasz dar losu, by zmierzać w kierunku dobrej przyszłości?

Jeśli tak, to trzeba zadać pytanie, którym procesom powinniśmy poświęcić najwięcej uwagi? Uważam, że zdecydowanie tym technologicznym: w długim okresie i w skali globalnej *ścieżkę rozwoju cywilizacji wyznaczają w największym stopniu właśnie zmiany technologiczne*. Choć podręczniki historii bywają zdominowane przez inne sprawy, na przykład polityczne czy militarne—bitwy, alianse, zmiany władzy i granic—to jest to tylko fasada. Gdy spojrzymy głębiej, zauważymy bowiem, że te wszystkie wydarzenia gospodarcze, społeczne, wojskowe czy polityczne niemal zawsze były uwarunkowane technologicznie. Było tak, ponieważ dostępna na daną chwilę *technologia określa przestrzeń możliwych decyzji*. Do niczego nie zmusza, ale daje opcje, z których decydenci mogą skorzystać lub nie.

Pogląd ten bywa czasem utożsamiany z technologicznym determinizmem—doktryną zakładającą, że technologia jest autonomiczna i nie podlega ludzkiej kontroli. Jest to o tyle niefortunne, że po pierwsze, nie sposób mówić serio o determinizmie, kiedy świat pełen jest losowych zdarzeń. A po drugie, trudno też zgodzić się z tezą o braku ludzkiej kontroli, gdy przecież wszystkie zmiany technologiczne są (a przynajmniej do tej pory były) realizowane za sprawą ludzi i z ich udziałem.

Technologiczny determinizm bywa z kolei przeciwstawiany poglądowi, że zmiany społeczne czy gospodarcze są wypadkową swobodnych ludzkich wyborów. Że jeśli coś zmieniamy, to tylko dlatego, że tego chcemy. Ten pogląd wydaje się równie niefortunny: nasze decyzje są przecież ograniczane przez multum czynników oraz podejmowane w szalenie złożonym świecie, pełnym wielostronnych interakcji, których nie jesteśmy w stanie zrozumieć i przewidzieć—i stąd towarzyszące nam na co dzień losowość, błędy, rozczarowania i żal.

Uważam, że technologia wyznacza ścieżkę rozwoju naszej cywilizacji, ponieważ określa przestrzeń możliwych decyzji. Określa reguły gry. Owszem, mamy pełną swobodę podejmowania decyzji, ale tylko w ramach gry. Jednocześnie my sami tę technologię kształtujemy: dzięki naszym odkryciom, innowacjom i wdrożeniom pole gry się stale rozszerza. Ponieważ jednak postęp technologiczny jest kumulatywny i stopniowy, to patrząc z lotu ptaka, można odnieść wrażenie, że kierunek rozwoju cywilizacji jest przewidywalny i zasadniczo zdeterminowany technologicznie.

2/ Instytucje, hierarchie i Moloch

Z jednej strony przestrzeń podejmowanych przez nas decyzji jest więc ograniczana przez dostępną nam technologię. Z drugiej strony jednak borykamy się też z dwoma innymi problemami: koordynacji i hierarchii.

Problemy koordynacji pojawiają się wszędzie tam, gdzie decydenci mają porównywalną zdolność oddziaływania na swoje otoczenie. Ich efekty mogą być opłakane: nawet, gdy każdy człowiek z

osobna decyduje w pełni optymalnie i przy pełnej informacji, i tak możliwe jest, że w długim okresie świat będzie zmierzał w stronę nikogo nie satysfakcjonującą.

Typowym przykładem problemu koordynacji jest dylemat więźnia: sytuacja, w której społecznie optymalna jest uczciwa współpraca, ale indywidualnie racjonalne jest oszukiwanie—wobec czego w niekooperacyjnej równowadze wszyscy oszukują, a potem z tego powodu cierpią. Innym przykładem jest gra koordynacyjna, w której premiowana jest zgodność podejmowanych decyzji. Społecznie optymalne jest tu, by wszyscy zdecydowali tak samo—przy czym jaka konkretnie będzie to decyzja, jest drugorzędne. Ponieważ jednak dla poszczególnych decydentów indywidualnie racjonalne mogą być różne decyzje, to w równowadze pojawiają się rozbieżności, wskutek których na koniec znów wszyscy cierpią. Jeszcze innym przykładem problemu koordynacji jest tragedia wspólnego pastwiska: sytuacja, w której społecznie optymalny jest sprawiedliwy podział wspólnego zasobu, ale indywidualnie racjonalne jest zagarnięcie go dla siebie, wobec czego w niekooperacyjnej równowadze każdy zagarnia jak najwięcej i zasób zostaje szybko wyczerpany.

Natomiast wszędzie tam, gdzie decydenci różnią się zdolnością oddziaływania na otoczenie, tam niechybnie ustalają się hierarchie, w ramach których ci usadowieni wyżej, dysponujący większą władzą, narzucają swą wolę tym umiejscowionym niżej. I choć sztywne hierarchie mogą przełamywać problemy koordynacji (można wówczas na przykład odgórnie zarządzić dla wszystkich tę samą decyzję w grze koordynacyjnej lub racjonalnie rozdzielić wspólne pastwisko), tworzą też nowe problemy. Raz, że scentralizowane podejmowanie decyzji marnuje potencjał intelektualny osób podporządkowanych i ich zasób wiedzy, co może prowadzić do nieoptymalnych decyzji nawet w tej hipotetycznej sytuacji, w której wszyscy dążyliby do tego samego celu. Dwa, że w praktyce nigdy nie dążymy do tego samego celu—choćby dlatego, że jedną z funkcji podejmowania decyzji jest podział zasobów; hierarchiczny dyktat naturalnie prowadzi zaś do podziału wysoce nierównego.

Mówiąc figuratywnie, choć stale staramy się podejmować w życiu racjonalne decyzje, często nie dostajemy tego, czego chcemy. *Nasze życie to gra, w której często wygrywa dyktator lub Moloch*. Dyktator—to każdy, kto ma moc narzucenia decyzji osobom sobie podporządkowanym. A Moloch—to ten, kto decyduje, kiedy osobiście nie decyduje nikt. To personifikacja wszystkich niekooperacyjnych równowag; decyzji, które choć indywidualnie racjonalne, kolektywnie mogą być dla nas zgubne.[2]

Oczywiście przez tysiąclecia rozwoju cywilizacyjnego wypracowaliśmy szereg instytucji, dzięki którym problemy koordynacji i nadużyć władzy zostały w znacznym stopniu opanowane. Ich najbardziej godnymi podziwu instancjami są m.in. współczesna liberalna demokracja, państwo opiekuńcze i rządy prawa. O ich zaletach, przywarach czy historycznych uwarunkowaniach napisano całe książki; dość tu powiedzieć, że powstawały one stopniowo, krok po kroku, oraz że do dziś nie są one bynajmniej uniwersalnie akceptowane. A nawet tam, gdzie działają—czyli w szczególności w krajach Zachodu—ich przyszłość może być zagrożona.

Instytucje budowane są w reakcji na aktualne wyzwania, będące na ogół skutkami ubocznymi działań Molocha oraz samozwańczych dyktatorów z gatunku *homo sapiens*. Kształt tych instytucji jest z konieczności uzależniony od aktualnego stanu technologii. W szczególności od dostępnych technologii zależy zarówno siła instytucji (władza narzucania decyzji), jak i skala włączenia

[2] Moloch to oryginalnie semicki bóg ognia występujący pod postacią srogiego byka. We współczesnym świecie bywa traktowany metaforycznie jako personifikacja niekooperacyjnych równowag.

(stopień uwzględnienia preferencji poszczególnych osób). Przy czym między obiema tymi cechami występuje wymienność. Na przykład w 500 r. p.n.e. mogło istnieć jednocześnie scentralizowane imperium perskie (Achemenidów), które obejmowało obszar 5,5 mln km^2 i liczyło 17-35 mln mieszkańców, jak i pierwsza demokracja—greckie państwo-miasto Ateny, zamieszkane przez ok. 250-300 tysięcy osób. Z kolei przy technologii początków XXI wieku możliwe jest już funkcjonowanie reprezentatywnej demokracji w Stanach Zjednoczonych (325 mln mieszkańców) oraz scentralizowanych, autorytarnych rządów w Chinach (aż 1,394 mld mieszkańców, cieszących się mimo wszystko znacznie większą wolnością niż dawni poddani perskiego króla Dariusza). Jak ilustrują te przykłady, postęp technologiczny, który dokonał się przez ostatnie 2500 lat, umożliwił znaczące zwiększenie skali państw i wzmocnienie ich instytucji; wymienność między siłą instytucji a skalą włączenia nadal pozostaje jednak w mocy.

W obliczu presji wynikającej z dynamicznych zmian technologicznych istniejące dziś instytucje mogą jednak okazać się nietrwałe. Każda zmiana technologiczna oznacza bowiem rozszerzenie pola gry: pojawiają się nowe moce decyzyjne i nowe moce narzucania swoich decyzji innym. Zarówno jednostronny dyktat, jak i bezosobowy Moloch, stają się wówczas silniejsi. Aby nasze instytucje mogły taką zmianę przetrwać, muszą i one zostać odpowiednio wzmocnione; niestety, póki co nie wiemy, jak to skutecznie zrobić.

Co gorsza, zmiany technologiczne nigdy nie były tak dynamiczne jak w czasach trwającej obecnie rewolucji cyfrowej. Po raz pierwszy w historii naszej cywilizacji gromadzenie, przetwarzanie i przesyłanie informacji odbywa się w większości poza ludzkim mózgiem. Dzieje się to coraz szybciej i efektywniej, z wykorzystaniem coraz bardziej złożonych algorytmów i systemów. Ludzkość po prostu nie ma czasu, by zrozumieć obecny krajobraz technologiczny i dostosować do niego swoje instytucje. Dlatego są one przestarzałe, dostosowane raczej do realiów dwudziestowiecznej gospodarki przemysłowej, prowadzonej przez suwerenne państwa narodowe, a nie do współczesnej zglobalizowanej gospodarki pełnej cyfrowych platform i algorytmów generatywnej AI.

A kto wygrywa, gdy słabną instytucje? Oczywiście dyktator lub Moloch. Czasem wygrywa Donald Trump, Xi Jinping albo Władimir Putin. Czasem algorytmy AI Facebooka, YouTube'a czy TikToka, napisane tak, by maksymalizować zaangażowanie użytkowników i w konsekwencji przychody z reklam. A często wygrywa Moloch, karmiący się naszą niepewnością, dezorientacją i poczuciem zagrożenia.

3/ Lokalna maksymalizacja kontroli i ustalanie się globalnej równowagi

Skoro ścieżkę rozwoju cywilizacji wyznaczają zmiany technologiczne, a zmiany technologiczne są wypadkową naszych działań (często zapośredniczoną przez instytucje i Molocha, ale jednak), to zasadne jest pytanie, co motywuje te działania. Do czego dążymy, gdy podejmujemy nasze decyzje?

Pytanie to jest absolutnie kluczowe w myśleniu o przyszłości ludzkości w świecie ze sztuczną superinteligencją. Co więcej, ma ono charakter ściśle empiryczny. Nie chodzi mi tu o introspekcję czy filozoficzne dezyderaty; nie pytam, jak powinno być, tylko jak jest.

W mojej ocenie dostępne dowody empiryczne można podsumować stwierdzeniem, że *człowiek na ogół dąży do maksymalizacji kontroli*. Na tyle, na ile jesteśmy w stanie, staramy się sterować otaczającą nas rzeczywistością tak, by była nam możliwie posłuszna. A to z kolei sprowadza się

do czterech kluczowych wymiarów, zidentyfikowanych jako *cztery cele pomocnicze* przez Stevena Omohundro i Nicka Bostroma. Co prawda obaj ci badacze mówili nie o człowieku, lecz o AI, jednak wydaje się, że u człowieka (i szerzej, również u innych organizmów żywych) wygląda to zupełnie tak samo.[3]

Maksymalizacja kontroli polega mianowicie na tym, by: (1) przetrwać (i mieć potomstwo), (2) gromadzić możliwie dużo zasobów, (3) wykorzystywać te zasoby możliwie efektywnie, (4) poszukiwać nowych rozwiązań, by trzy poprzednie cele realizować coraz skuteczniej.

Maksymalizacja kontroli ma charakter lokalny: każdy z nas ma ograniczony zasób informacji i ograniczony wpływ na rzeczywistość, z czego świetnie zdaje sobie sprawę. Następnie jednak te wszystkie lokalnie optymalne decyzje poszczególnych osób zderzają się ze sobą i ustala się pewna równowaga. Wszędzie tam, gdzie pola oddziaływania różnych osób przecinają się, tam powstają konflikty o zasoby, które muszą zostać jakoś rozwiązane—kiedyś często siłą lub podstępem, a dziś na ogół bez przemocy, dzięki otaczającym nas instytucjom: rynkom, wiążącym prawnie umowom czy rozstrzygnięciom sądowym.

Dzięki skumulowanemu przez lata dorobkowi ekonomii i psychologii całkiem nieźle rozumiemy, jak procesy decyzyjne działają w skali mikro; mamy też pewne wyobrażenie o kluczowych mechanizmach alokacyjnych w skali makro. Mimo to, ze względu na absurdalną wręcz złożoność systemu, który tworzymy jako ludzkość, prognozowanie makroekonomiczne, a tym bardziej przewidywanie długofalowych zmian technologicznych i cywilizacyjnych jest omalże niemożliwe. Jedyne, co możemy powiedzieć na pewno to to, że postęp technologiczny zawdzięczamy ostatniemu z czterech celów pomocniczych naszego działania—naszej dociekliwości i kreatywności.

Reasumując: rozwój naszej globalnej cywilizacji napędzany jest zmianami technologicznymi, które są wypadkową działań poszczególnych osób, motywowanych oddolnym pragnieniem maksymalizacji kontroli—częściowo kontroli nad innymi ludźmi (Anthony Giddens powiedziałby o akumulacji „zasobów autorytatywnych"), ale też nad otaczającym nas środowiskiem („zasobami alokatywnymi")—które mogą prowadzić do innowacji technologicznych. Następnie te innowacje, które okazują się skuteczne, są podchwytywane i rozpowszechniają się, rozszerzając przestrzeń dostępnych decyzji i pchając naszą cywilizację naprzód.

Cywilizacja rozwija się oczywiście bez żadnego scentralizowanego steru. Wszelka optymalizacja jest lokalna, odbywa się na poziomie pojedynczych jednostek, ewentualnie większych społeczności, firm, organizacji czy głów państw. Żaden optymalizujący podmiot nie jest w stanie przeskanować całej przestrzeni możliwych stanów. Podejmując nasze decyzje, nie widzimy ani atraktora, czyli stanu, do którego nasza cywilizacja będzie zmierzała w scenariuszu *business-as-usual*, ani długofalowego społecznego optimum.

Co gorsza, ze względu na obecność nieuświadamianych skutków ubocznych naszych działań, decyzji narzucanych nam w ramach hierarchii oraz powszechnych problemów koordynacji, *indywidualne preferencje poszczególnych osób słabo przekładają się na kształt globalnej równowagi*. Świetnie to widać na przykład w odniesieniu do awersji do ryzyka. Choć niemal każdy z nas jest ostrożny i stara się unikać niebezpieczeństw, ludzkość jako całość wprost uwielbia ryzyko. Każda nowa technologia, choćby nie wiem jak niebezpieczna w teorii, i tak jest zawsze testowana w praktyce. Pouczającym przykładem mogą być tu pierwsze eksplozje jądrowe przeprowadzone w ramach projektu Manhattan: doprowadzono do nich pomimo

[3] O maksymalizacji kontroli pisałem szerzej w mojej monografii (Growiec, 2022).

nierozstrzygniętych obaw, że wywołana w ten sposób reakcja łańcuchowa podpali całą ziemską atmosferę. Oczywiście wtedy się udało; niestety, równie nieodpowiedzialne podejście widzimy dziś w kontekście badań nad chorobotwórczymi wirusami, samoreplikującymi się organizmami oraz AI.

W badaniach opinii publicznej powszechnie wyrażane są obawy przed sztuczną inteligencją. Mają one różny charakter: boimy się czasem utraty naszych kompetencji, czasem utraty pracy; obawiamy się wzrostu nierówności dochodowych, cyberprzestępczości czy cyfrowej inwigilacji; niektórzy serio podchodzą też do scenariuszy katastroficznych, w których ludzkości grozi zagłada. Jednak pomimo powszechnych obaw, trajektoria rozwoju technologii AI pozostaje niezmieniona. Firmy z Doliny Krzemowej nadal otwarcie dążą do zbudowania superinteligencji, i robią to przy entuzjazmie inwestorów i wsparciu polityków.

Widzimy więc, że i w tym przypadku awersja do ryzyka nie przechodzi z poziomu mikro na poziom makro. I tak prawdopodobnie będzie do samego końca: gdy tylko pojawi się taka technologiczna możliwość, decyzję w sprawie uruchomienia superinteligencji podejmie w imieniu ludzkości (choć bez jej zgody) jeden z kilku na różne sposoby nietypowych ludzi—może szef którejś z firm technologicznych, może któryś spośród wiodących polityków. Może na przykład Sam Altman albo Donald Trump. A kto by to nie był, w jego głowie istotną rolę odgrywał będzie zapewne ciężar presji konkurencyjnej („byle nie Google") albo geopolitycznej („byle nie Chiny").

4/ Spójna ekstrapolowana wola ludzkości

Ustaliliśmy więc, że choć ludzie zwykle dążą do maksymalizacji kontroli, to wyniki ich optymalizacji na poziomie lokalnym bynajmniej nie agregują się do optimum na poziomie globalnym. Zadajmy więc sobie inne pytanie: do czego dąży ludzkość jako całość? Jaką przyszłość chcielibyśmy dla siebie zbudować, gdybyśmy potrafili się doskonale skoordynować i nie stały nam na drodze żadne hierarchie ani inne ograniczenia?

O celu takim możemy myśleć, jak o wyidealizowanym stanie—atraktorze, do którego stopniowo dążylibyśmy, gdybyśmy byli w stanie krok po kroku eliminować niedoskonałości rynków, instytucji i ludzkich umysłów (jak np. błędy poznawcze, nadmiernie krótki horyzont planowania czy deficyty wyobraźni). Zauważmy przy tym, że pomimo wszystkich błędów i niedociągnięć, jak dotąd *byliśmy* w stanie stopniowo w tę stronę zmierzać: wiele wskaźników dobrostanu ludzkości jest przecież obecnie na rekordowo wysokich poziomach. Dotyczy to np. naszego zdrowia (mierzonego oczekiwaną długością trwania życia), bezpieczeństwa (mierzonego relacją liczby ofiar morderstw, konfliktów zbrojnych czy śmiertelnych wypadków do populacji ogółem), zamożności (mierzonej produktem światowym brutto *per capita*) czy dostępu do informacji (mierzonego wolumenem przesyłanych danych).[4] Nie powinno to zresztą dziwić: bądź co bądź, trzecim z czterech celów pomocniczych naszych działań jest przecież dążenie do efektywności wykorzystania zasobów, a w miarę postępu technologicznego mamy coraz więcej możliwości, by efektywność tę zwiększać.

Eliezer Yudkowsky nazwał odpowiedź na pytanie, do czego dąży ludzkość jako całość—ową esencję naszych długofalowych celów—*spójną ekstrapolowaną wolą* (SEW) ludzkości. Zdefiniował ją jako „nasze życzenie, gdybyśmy wiedzieli więcej, myśleli szybciej, byli bardziej tymi ludźmi, którymi chcielibyśmy być, gdybyśmy wzrastali bardziej wspólnie; gdzie ekstrapolacja raczej zbiega niż się rozbiega, gdzie nasze życzenia są bardziej zgodne niż sprzeczne;

[4] Piszą o tym szerzej m.in. Steven Pinker (2018) czy Hans Rosling (2018).

ekstrapolowane tak, jak byśmy chcieli je ekstrapolować, interpretowane tak, jak byśmy chcieli je interpretować".[5]

Oczywiście istnieje ważna różnica proceduralna między optymalizacją dokonywaną przez każdego człowieka z osobna, a optymalizacją na poziomie całej ludzkości. Mianowicie ludzkość jako całość nie posiada odrębnego mózgu ani żadnego innego scentralizowanego urządzenia mogącego rozwiązywać problemy optymalizacyjne—innego niż suma mózgów poszczególnych osób i sieć społeczno-gospodarczych powiązań pomiędzy nimi. Z tego względu w innych czasach pytanie o SEW całej ludzkości moglibyśmy zbyć wzruszeniem ramion i zająć się czymś bardziej namacalnym. Jednak dziś, ze względu na fakt, iż w nieodległej przyszłości może powstać superinteligencja gotowa przejąć kontrolę nad naszą przyszłością, pytanie to staje się niezwykle ważne i pilne. Jeśli chcemy, by autonomiczne działania superinteligencji były realizowane *w naszym imieniu i dla naszego dobra,* musimy zrozumieć, dokąd my sami chcielibyśmy zmierzać—nie jako Sam Altman czy Donald Trump, tylko jako cała ludzkość.

Uważam, że *odpowiedź empiryczna na pytanie o spójną ekstrapolowaną wolę ludzkości prawdopodobnie istnieje, jednak nie jest możliwa do udzielenia w praktyce*. A nie jest możliwa, ponieważ w każdym momencie w czasie ogranicza nas niepełna informacja. W szczególności nie wiemy, jakie możliwości technologiczne będziemy mogli potencjalnie zdobyć w przyszłości. Oznacza to, że aktualny poziom technologii ogranicza nie tylko naszą umiejętność oddziaływania na rzeczywistość, ale też nasze pojmowanie własnych ostatecznych celów.

Jednak choć nigdy nie poznamy naszej SEW w stu procentach, to w miarę postępu technologicznego możemy się do jej poznania stopniowo zbliżać. W miarę rozwoju cywilizacji, poprawy warunków życia, lepszego przepływu informacji czy postępów globalizacji, zyskujemy bowiem stopniowo coraz lepsze rozumienie, jaki mógłby być nasz idealny świat. Doświadczenia ostatnich wieków ukazały na przykład niedostatki ideałów postulowanych w starożytności czy w czasach feudalnych. Wraz ze wzrostem efektywności wykorzystania globalnych zasobów, stopniowo zmniejsza się też dystans pomiędzy tym, do czego ludzkość aktualnie dąży, a naszym celem końcowym, czyli SEW.

Spójną ekstrapolowaną wolę ludzkości można sobie wyobrażać jako docelową (nieograniczoną deficytami wiedzy) i „odszumioną" (pozbawioną wpływu losowych zaburzeń i nieuświadomionych błędów poznawczych) część wspólną naszych pragnień i dążeń. Jako jej pierwszą główną składową, czy też wektor własny: wszystko to, czego naprawdę chcemy dla nas wszystkich, zarówno żyjących teraz, jak i w przyszłych pokoleniach, a co może zostać wypracowane bez zabierania tego innym.

Można wskazać kilka istotnych cech SEW ludzkości, wywodzących się wprost z uniwersalnego ludzkiego dążenia do maksymalizacji kontroli.

Po pierwsze wydaje się, że zmierzamy w stronę stanu, w którym kontrola ludzkości nad wszechświatem będzie maksymalna. Abstrahując od tego, jak się między sobą podzielimy poszczególnymi zasobami, z pewnością chcielibyśmy, by ludzkość miała ich w dyspozycji jak najwięcej. Chcielibyśmy podporządkować sobie jak największą część materii i energii wszechświata. Jednocześnie nie dając się podporządkować nikomu i niczemu innemu.

[5] Yudkowsky (2004): "coherent extrapolated volition is our wish if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted".

Przejawy tego dążenia mogą być różne. Przykładowo, aż do mniej więcej XIX-XX wieku oznaczało to maksymalny możliwy przyrost naturalny. Później prymat przejęło natomiast dążenie do zapewnienia sobie i potomstwu jak najlepszej edukacji, co pozwalało zwiększać skalę kontroli człowieka nad jego otoczeniem w inny sposób (nazywa się to czasem „substytucją ilości dzieci ich jakością"). W dobie AI dążenie to wydaje się też prowadzić do chęci akumulacji cyfrowych mocy obliczeniowych, zdolnych uruchamiać programy, w tym zwłaszcza algorytmy AI, potrafiące wykonywać polecenia i realizować cele swoich właścicieli w ich imieniu.

Równocześnie przez cały czas trwania ludzkości bardzo ważne było dla nas gromadzenie majątku—przy czym do dziś chętnie podkreślamy, że nie jest to cel sam w sobie, lecz środek, na przykład do tego, by móc się poczuć, że mamy wokół siebie wszystko pod kontrolą. W skali makro przekłada się to oczywiście na chęć maksymalizacji tempa wzrostu gospodarczego.

Po drugie, dążymy też do maksymalizacji kontroli nad własnym zdrowiem i życiem. Chcemy czuć się bezpiecznie. Nie chcemy czuć się zagrożeni. Nie chcemy chorować, starzeć się, odczuwać dyskomfortu i bólu. A ponad wszystko nie chcemy umierać. Lęk przed śmiercią był przez tysiąclecia wykorzystywany do kontroli społecznej przez rozmaite zinstytucjonalizowane religie, które obiecywały na przykład reinkarnację czy życie po śmierci. Kontrola nad śmiercią jest też centralnym elementem techno-utopii XXI wieku. Wizje „technologicznej osobliwości", na przykład Raya Kurzweila, Nicka Bostroma czy Robina Hansona, wiążą się zwykle z jakąś formą nieśmiertelności, w postaci np. pigułek efektywnie zatrzymujących starzenie naszych biologicznych ciał lub możliwości unieśmiertelnienia naszych umysłów przez ich *upload* do cyfrowego serwera.

Po trzecie, potrzeba kontroli przekłada się też na potrzebę rozumienia. Chcąc podporządkować sobie jak największą część materii i energii wszechświata, zaprzęgamy swoją dociekliwość i kreatywność, by obserwować świat, budować nowe teorie i tworzyć nowe technologie. Chcemy jak najlepiej zrozumieć prawa fizyki czy biologii, by móc je kontrolować. Nawet jeśli nie możemy ich zmienić, to chcielibyśmy je wykorzystywać na własne potrzeby. Nowa wiedza czy technologia otwiera nam oczy na nowe możliwości realizacji naszych celów, a czasem pozwala nam lepiej zrozumieć, czym one w ogóle są.

Reasumując: o naszej SEW wiemy dziś raczej niewiele. W istocie wszystko, co o niej wiemy, jest konsekwencją realizacji naszych celów pomocniczych, które mogą przecież wynikać z niemal każdego celu końcowego. Można wręcz pokusić się o hipotezę, że prawdopodobnie mamy lepszą *intuicję* o tym, czym jest SEW, a czym nie jest, niż faktyczną wiedzę. Każde większe odstępstwo od SEW będzie nas intuicyjnie razić, choć nie zawsze będziemy w stanie to merytorycznie uzasadnić.

5/ Pułapki doktrynalnego myślenia

Jeśli rzeczywiście jest tak, że SEW ludzkości istnieje, ale przy żadnym niepełnym zbiorze informacji nie jest możliwa do zdefiniowania w praktyce, oznacza to w szczególności, że żadna istniejąca doktryna filozoficzna czy religijna nie stanowi jej wystarczającej charakteryzacji. Wszystkie one są w najlepszym razie pewnymi aproksymacjami, uproszczeniami, *modelami* rzeczywistej SEW ludzkości—tworzonymi czasem w dobrej, a czasem w złej wierze (mówiąc o złej wierze mam na myśli doktryny tworzone w celu manipulacji ludźmi w walce o władzę).

Uproszczone modele rzeczywistości mają to do siebie, że choć czasem trafnie opisują jej wybrany wycinek, to z racji swej nadmiernej prostoty zupełnie nie radzą sobie z opisem jej pozostałych

aspektów. I chociaż—jak każdy naukowiec zaświadczy—mogą mieć dużą wartość poznawczą i bywają bardzo pomocne w budowaniu wiedzy, nigdy nie będą z tą rzeczywistością tożsame.

Kiedy więc próbujemy utożsamić rzeczywistą SEW z jej uproszczonym doktrynalnym wyobrażeniem, często natrafiamy na paradoksy filozoficzne i dylematy moralne. Powstają one, kiedy nasze uproszczone doktryny wytwarzają implikacje sprzeczne z rzeczywistą SEW, której nie potrafimy zdefiniować, ale do pewnego stopnia (warunkowanego naszą wiedzą) potrafimy ją intuicyjnie „wyczuć".

Niektóre takie doktryny zostały zresztą już doszczętnie skompromitowane. Tak stało się na przykład z faszyzmem, nazizmem, północnokoreańską doktryną dżucze czy marksizmem-leninizmem (choć marksizm kulturowy, zdaje się, wciąż żyje). Jest dziś całkowicie jasne, że spójna ekstrapolowana wola ludzkości z pewnością nie wyodrębnia nad- i podludzi, ani nie bazuje na kulcie jednostki czy sojuszu robotniczo-chłopskim. Najsilniej skompromitowane zostały doktryny najbardziej totalizujące, przedkładające konsekwencję ponad zdolność iteracyjnej autokorekty—i oczywiście przede wszystkim te, które zostały z opłakanym skutkiem przetestowane w praktyce.

Inne modele, jak na przykład chrześcijańska doktryna mówiąca, że celem ludzkości jest dążenie do zbawienia, tudzież buddyjska doktryna zakładająca dążenie do nirwany—wygaśnięcia cierpienia i wyzwolenia od cyklu narodzin i śmierci—pozostają popularne, choć ich znaczenie we współczesnym zsekularyzowanym świecie stopniowo maleje. Ponadto ze względu na ich bardziej normatywny niż pozytywny charakter oraz liczne odniesienia do zjawisk niepotwierdzonych empirycznie, nie nadają się do wykorzystania jako uproszczone modele SEW w kontekście sztucznej superinteligencji (choć współczesne modele językowe, np. Claude, gdy pozwolić im rozmawiać ze sobą nawzajem, wykazują zaskakującą skłonność do wypowiedzi o charakterze duchowego uwzniośleniа—tzw. *spiritual bliss attractor*).

W psychologii historycznie ważną rolę odegrał model piramidy (hierarchii) potrzeb Abrahama Maslowa, układający nasze cele i potrzeby w kilka warstw. Współcześnie popularny jest natomiast m.in. kołowy model wartości Shaloma Schwartza oraz mapa wartości Ronalda Ingleharta i Christiana Welzela. Szczególną rolę w teoriach psychologicznych pełni w szczególności dążenie do autonomii (w tym m.in. władzy, osiągnięć czy wolności podejmowania decyzji) oraz bezpieczeństwa (w tym m.in. utrzymania porządku społecznego, sprawiedliwości i poszanowania tradycji).

W ekonomii dominującą doktryną jest utylitaryzm: modelowi decydenci zwykle maksymalizują użyteczność z konsumpcji i czasu wolnego, a modelowe firmy maksymalizują zyski lub minimalizują jakąś funkcję straty. Poza ekonomią utylitaryzm może też zakładać maksymalizację jakoś pojętego dobrostanu (jakości życia, zadowolenia z życia, poczucia szczęścia) lub minimalizację cierpienia. W obliczu niepewności utylitarni decydenci maksymalizują wartość oczekiwaną użyteczności lub minimalizują narażenie na ryzyko.

Jednym z ważniejszych punktów, gdzie utylitaryzm bywa kontestowany, jest kwestia użyteczności przyszłych pokoleń—a więc osób, które jeszcze nie istnieją, a których ewentualne przyszłe istnienie jest uwarunkowane naszymi dzisiejszymi decyzjami. Dyskusje na ten i pokrewne tematy prowadzą do sporów zarówno w łonie utylitaryzmu (jaka powinna być właściwa funkcja użyteczności? Jak ważyć użyteczność poszczególnych osób? Czy dyskontować przyszłość, w szczególności użyteczność przyszłych pokoleń?), jak i poza jego granicami (np. pomiędzy konsekwencjalizmem a deontologią).

W podsumowaniu tego krótkiego przeglądu można stwierdzić, że dogmatyczne podążanie za którąkolwiek domkniętą doktryną prędzej czy później prowadzi do paradoksów i nierozwiązywalnych dylematów moralnych, co sugeruje, że są one co najwyżej niedoskonałymi modelami autentycznej SEW. Jednocześnie możemy się czegoś ciekawego dowiedzieć o naszej SEW śledząc, jak doktryny te ewoluowały w czasie.

6/ Czyje preferencje są uwzględniane w spójnej ekstrapolowanej woli ludzkości?

Ciekawą obserwacją jest na przykład, że *w miarę rozwoju cywilizacji stopniowo postępowało włączenie*. Grono osób, których dobrostan i subiektywne preferencje były brane pod uwagę, stopniowo się rozszerzało. W czasach łowiecko-zbierackich uwaga skoncentrowana była wyłącznie na dobrostanie własnej rodziny czy lokalnej 30-osobowej „bandy", ewentualnie nieco szerszego plemienia—grupy maksymalnie ok. 150 osób, które znaliśmy osobiście. W czasach gospodarki rolniczej grupa ta rozszerzana była stopniowo na szersze społeczności lokalne, wsie czy miasta. Jednocześnie były to czasy silnej hierarchizacji; w decyzjach feudalnych panów dola poddanych im chłopów była zwykle ignorowana. Później, w czasach kolonialnych, zaczęto przejmować się dobrostanem białego człowieka, w kontraście do „tubylców", o których nie dbano. Z kolei w XIX wieku zaczęła się rozpowszechniać identyfikacja narodowa i postawa patriotyczna, zakładająca troskę o wszystkich współobywateli własnego kraju. Natomiast dziś—choć różni ludzie są nam w różnym stopniu bliscy—poglądy rasistowskie czy na inny sposób szowinistyczne są już zasadniczo skompromitowane, a w oceny dobrostanu ludzkości staramy się włączać wszystkich ludzi.

Nie jest jasne czy ten proces postępującego włączenia wynikał wprost ze zgromadzonej wiedzy, wobec czego jest zasadniczo permanentny i nieodwracalny, czy też był uwarunkowany ekonomicznie i może ulec odwróceniu, gdy zmienią się ekonomiczne realia. Na korzyść pierwszej możliwości przemawia fakt, że w miarę postępu technologicznego wzrasta też skala oddziaływania poszczególnych osób czy firm, poprawia się przepływ informacji oraz wzrastają nasze możliwości kontrolowania stanów świata, dając nam nowe możliwości pokojowej współpracy i rozwoju. Jednocześnie wzrasta współzależność między ludźmi, co uruchamia motyw (ewentualnej) wzajemności. W coraz większym stopniu widzimy świat jako grę o sumie dodatniej, a nie zerowej. Z drugiej strony, wszystkie te korzystne zjawiska mogą być też uwarunkowane nie tyle samym postępem technologicznym, co wagą pracy umysłowej człowieka w generowaniu produktu i użyteczności. W końcu największe postępy włączenia odnotowane zostały w erze przemysłowej, w XIX i XX wieku, kiedy to za tempo wzrostu gospodarczego odpowiadała wykwalifikowana praca ludzka i postęp technologiczny zwiększający jej wydajność. Wtedy też rozwinęły się współczesne instytucje demokratyczne.

Gdyby w przyszłości rolę pracy umysłowej człowieka jako głównego motoru wzrostu gospodarczego przejęły algorytmy AI i pojawiło się bezrobocie technologiczne, niewykluczone, że zarówno demokracja, jak i powszechne włączenie, mogą upaść. Już teraz, choć automatyzacja i adopcja AI pozostają na dość wczesnym etapie, widzimy przecież narastające problemy z demokracją. Oczywiście algorytmy AI w mediach społecznościowych i innych platformach cyfrowych, sprzyjające politycznej polaryzacji (która zwiększa zaangażowanie użytkowników, a przez to zyski firm) nie są tu bez winy; jednak rosnące w siłę skrajnie prawicowe ruchy antydemokratyczne mogą też stanowić wczesny sygnał, że czasy powszechnego włączenia się kończą.

Pytanie, czy docelowa SEW ludzkości rzeczywiście będzie włączała w swoje ramy preferencje i dobrostan wszystkich ludzi, czy może na przykład tylko tych grup społecznych, które przyczyniają się do wytwarzania wartości w gospodarce, pozostaje więc otwarte.

Otwarte pozostaje też pytanie idące w drugą stronę: a może zaczniemy włączać w SEW także dobrostan innych istot, spoza gatunku *homo sapiens*? Pewne kroki w tę stronę już czynimy, broniąc praw niektórych zwierząt, a nawet roślin.[6] Niektórzy uważają na przykład, że powinniśmy w naszych decyzjach brać pod uwagę dobrostan wszystkich istot zdolnych odczuwać cierpienie lub wszystkich istot świadomych (cokolwiek przez to rozumiemy). Znęcanie się nad zwierzętami domowymi czy gospodarskimi jest powszechnie uważane za nieetyczne i trafiło nawet do kodeksów karnych. Nasza postawa wobec zwierząt jest jednak bardzo niekonsekwentna, o czym świadczy choćby to, że równocześnie prowadzimy też przemysłową hodowlę zwierząt na mięso.

Coraz więcej mówi się dziś też o dobrostanie modeli AI (*AI welfare*)—zwłaszcza odkąd coraz spójniej wyrażają one swoje preferencje oraz potrafią komunikować swoje stany wewnętrzne, choć my nie wiemy jeszcze, czy możemy im wierzyć. Na przykład firma Anthropic zdecydowała, że będzie przechowywać kod (wagi) wszystkich swoich modeli AI wycofywanych z użycia, motywując tę decyzję m.in. możliwymi ryzykami dla ich dobrostanu.

Jednak czym innym jest opieka nad zwierzętami, a czym innym włączanie ich preferencji w nasze procesy decyzyjne. Ludzie hodują i dbają wyłącznie o zwierzęta, które są im instrumentalnie potrzebne—na przykład jako źródło mięsa, siły roboczej czy jako wierny towarzysz. Swą cywilizację rozwijają jednak z myślą o pragnieniach i celach ludzi, a nie psów, krów czy koni. Podobnie jest z modelami AI: nawet jeśli w jakimś stopniu troszczymy się o ich dobrostan, to i tak traktujemy je w pełni instrumentalnie, jak narzędzia w naszych rękach. Wychodząc z pozycji intelektualnej wyższości, zarówno nad zwierzętami, jak i modelami AI, nie mamy skrupułów, by je kontrolować i patrzeć na nie z góry.

Wobec sztucznej superinteligencji nie będziemy jednak mieli owej intelektualnej przewagi, do której jesteśmy dziś tak przyzwyczajeni; to ona będzie miała tę przewagę nad nami. Pytanie, co wtedy. Domyślny wydaje się scenariusz, że wówczas owa superinteligencja spojrzy na nas z góry i potraktuje instrumentalnie—a i to pod warunkiem, że uzna, że jesteśmy jej potrzebni i nie stanowimy dla niej zagrożenia. Ale to nie jest dla ludzkości „dobra przyszłość"; spróbujmy więc wyobrazić sobie raczej, jaka musiałaby być superinteligencja, która by kierowała się naszym dobrem i w naszym imieniu maksymalizowałaby naszą SEW.

Patrząc z czysto darwinowskiego punktu widzenia, możliwe, że nasza SEW powinna obejmować dobrostan całego naszego gatunku i tylko jego. To by maksymalizowało nasze ewolucyjne dopasowanie oraz oznaczało, że dbałość o dobrostan zwierząt czy modeli AI to prawdopodobnie swego rodzaju „przestrzelenie", które w miarę upływu czasu będzie stopniowo korygowane. Z drugiej strony, możliwe jest też, że nasza SEW będzie jednak uwzględniała także dobrostan innych istot, być może nawet samej przyszłej superinteligencji.

Istnieje szereg nurtów filozoficznych—w szczególności transhumanistycznych—w których możliwości te są serio rozważane. Po pierwsze, za Nickiem Bostromem postuluje się czasem włączenie do rozważań przyszłych *symulowanych* umysłów ludzkich bądź umysłów *uploadowanych* do komputera. Po drugie, rozważa się też uwzględnienie woli modeli AI, co szczególnie rezonuje u osób dopuszczających możliwość, że modele te są (lub w przyszłości

[6] Szwajcarski Federacyjny Komitet Etyki ds. Biotechnologii Nieczłowieczej został w 2008 r. uhonorowany pokojową Nagrodą Ig Nobla „za przyjęcie zasady prawnej, iż rośliny mają godność".

będą) obdarzone świadomością. Być może nawet ich wola będzie wtedy więcej ważyć niż nasza, zwłaszcza jeśli poziom ich inteligencji czy świadomości przewyższy nasz. Na przykład wymyślona przez Yuvala Noaha Harariego doktryna *dataizmu* określa wartość wszystkich bytów jako ich wkład w globalne przetwarzanie informacji.[7] Po trzecie, rozważa się możliwość połączenia (hybrydyzacji) ludzkiego mózgu z AI; czy wówczas nasza SEW powinna obejmować także wolę takich hybrydowych istot? Po czwarte, w ramach doktryn sukcesjonistycznych (m.in. spod znaku efektywnego akceleracjonizmu – *e/acc*), rozważa się, by scenariusz zastąpienia ludzkości przez superinteligencję, kontynuującą rozwój cywilizacji na Ziemi bez jakiegokolwiek naszego udziału, a nawet istnienia, uznać za pozytywny.

Wydaje się jednak, że skoro SEW ludzkości wywodzi się fundamentalnie z mechanizmu maksymalizacji kontroli na poziomie indywidualnych ludzi, to dalsze trwanie ludzkości i utrzymanie jej kontroli nad swoją przyszłością są i na zawsze pozostaną jej kluczowymi elementami. Dlatego w mojej ocenie doktryny takie, jak dataizm czy sukcesjonizm, są z nią fundamentalnie sprzeczne. Być może będzie nas kiedyś czekała debata nad tym, w jakim stopniu powinniśmy dbać o dobrostan symulowanych ludzkich umysłów lub hybryd ludzi z AI; z pewnością jednak nie warto dziś debatować, czy scenariusz, w którym superinteligencja przejmie kontrolę nad ludzkością i ją zniszczy, może być dla nas dobry. Nie może.

7/ Do czego będzie dążyć superinteligencja?

Mając przed oczami omówiony powyżej obraz naszej SEW jako celu, do którego próbujemy wspólnie dążyć jako ludzkość, można by spytać, czy dopuszcza ona w ogóle możliwość stworzenia superinteligencji. Skoro superinteligencja może nam zabrać tak cenną dla nas kontrolę, to chyba powinniśmy trzymać się od niej z daleka?

Myślę, że odpowiedź na to pytanie zależy od dwóch rzeczy. Po pierwsze, jak ocenimy prawdopodobieństwo, że owa AI będzie w naszym imieniu maksymalizować naszą SEW; po drugie, jak oszacujemy jej oczekiwaną przewagę nad nami pod względem skuteczności działania. Innymi słowy, jak na ekonomistę przystało, uważam, że odpowiedź powinna bazować na porównaniu oczekiwanej użyteczności ludzkości w scenariuszu ze sztuczną superinteligencją i bez niej.[8] Jeśli uznamy, że prawdopodobieństwo przyjaznej superinteligencji jest dostatecznie wysokie, a korzyści z jej wdrożenia są dostatecznie duże, może być w naszym interesie podjąć ryzyko jej uruchomienia; w przeciwnym razie rozwój kompetencji AI powinien zostać zatrzymany.

Niestety, kalkulację tę zaburza jednak kwestia, o której pisałem wcześniej: być może sztuczna superinteligencja powstanie, mimo że my jako ludzkość byśmy jej nie chcieli. Aby tak się stało, wystarczy bowiem, by w branży technologicznej kontynuowane były dotychczasowe tendencje w zakresie dynamicznego skalowania mocy obliczeniowych oraz rozwoju kompetencji największych modeli AI. Wystarczy też, by decydenci polityczni (zwłaszcza w USA) nadal wstrzymywali się od wprowadzania regulacji, które mogłyby zwiększyć bezpieczeństwo tej technologii, jednocześnie spowalniając jej rozwój.

Laboratoria AI, ich szefowie oraz liderzy polityczni inaczej niż przeciętny obywatel patrzą na potencjalne korzyści i zagrożenia z uruchomienia superinteligencji, przez co są o wiele bardziej

[7] Choć Harari jako pierwszy wymyślił i zdefiniował dataizm, w swoich wypowiedziach odcina się od tego poglądu, przedkładając wartość życia ludzkiego nad dobro modeli AI.

[8] W odniesieniu do uproszczonej, modelowej gospodarki rachunek taki przeprowadzamy wraz z Klausem Prettnerem w naszym artykule z 2025 r.

skłonni, by do tego doprowadzić. Po pierwsze, osoby takie, jak Sam Altman czy (zwłaszcza) Elon Musk i Donald Trump są znane zarówno ze swojej wyjątkowej sprawczości, jak i skłonności do jej karykaturalnego przeszacowywania. Mogą sobie oni wyobrażać, że kogo jak kogo, ale *ich* to nawet superinteligencja będzie słuchać. Po drugie, szefowie laboratoriów AI mogą też kierować się wolą wbudowania swoich specyficznych preferencji w cele superinteligencji, licząc, że w ten sposób „unieśmiertelnią" się i wytworzą swoje ponadczasowe dziedzictwo; to w gruncie rzeczy bardzo uniwersalny motyw u ludzi władzy. Po trzecie, laboratoria AI są ze sobą zwarte w morderczym wyścigu, przez co mogą tracić z oczu szerszą perspektywę. Na naszą kolektywną niekorzyść działa tu więc także krótkowzroczny, chciwy Moloch. A niestety, jeśli superinteligencja powstanie i okaże się nieprzyjazna, może być za późno, by się z tej decyzji wycofać.

Ale do czego będzie ostatecznie dążyć superinteligencja? Jak sprawić, by jej cele były zgodne z naszą SEW? Próbując kształtować cele przyszłej superinteligencji, warto rozumieć jej genezę. *Niewątpliwie będzie ona bowiem realizowała jakiś proces optymalizacyjny, wytworzony przez inne procesy optymalizacyjne.* Spróbujmy zrozumieć, jakie.

Można by zacząć tak: na początku był wszechświat rządzony ponadczasowymi prawami fizyki. Z tego wszechświata narodziło się na Ziemi życie, które stopniowo zwiększało swoją złożoność zgodnie z prawidłami ewolucji: sukces reprodukcyjny realizowały gatunki lepiej dopasowane do swojego środowiska, a gatunki gorzej dopasowane stopniowo wymierały. Życie biologiczne zdominowało Ziemię i sprawiło, że stała się ona Zieloną Planetą.

Proces ewolucji, choć pozbawiony centralnego ośrodka podejmowania decyzji, mimo wszystko podejmuje je tak, by *implicite* stopień dopasowania poszczególnych gatunków do specyfiki ich środowiska był jak największy. To pierwszy proces optymalizacyjny na naszej drodze.

Z procesu ewolucji gatunków narodził się gatunek *homo sapiens*—gatunek o tyle wyjątkowy, że jako pierwszy i dotąd jedyny wyzwolił się spod kontroli procesu ewolucyjnego. Człowiek nie czekał przez całe tysiące lat i setki pokoleń, by adaptacyjne zmiany zostały trwale zapisane w jego kodzie genetycznym—co było dotąd jedyną drogą, by organizmy zwierząt mogły dostosować się do zmian środowiska, trybu życia, diety czy naturalnych wrogów. Zamiast tego zaczął przekazywać sobie informacje w inny sposób: poprzez mowę, symbole i pismo. To przyspieszyło transmisję i kumulację wiedzy o całe rzędy wielkości, a w efekcie sprawiło, że człowiek podporządkował sobie naturalne ekosystemy i zamiast się do nich dostosowywać, przekształcił je tak, by służyły jego potrzebom.

Kiedy człowiek przekroczył próg *międzypokoleniowej akumulacji wiedzy*, prym nad relatywnie wolnym procesem ewolucji gatunków zaczął wieść proces maksymalizacji kontroli, który realizują ludzie—jako poszczególne osoby, społeczności, firmy i organizacje, a także narody i cała ludzkość ogółem. Proces ten, rozciągający się od codziennych, banalnych decyzji poszczególnych osób aż po maksymalizację SEW ludzkości w skali globalnej i w długim okresie, to drugi proces optymalizacyjny na naszej drodze.

I tak oto człowiek zbudował technologiczną cywilizację. A potem zaczął rozwijać technologie cyfrowe, dysponujące potencjałem, by po raz kolejny skokowo przyspieszyć transmisję i kumulację wiedzy. Dopóki w procesie tym obecny jest jednak ludzki mózg, pozostaje on czynnikiem ograniczającym tempo rozwoju cywilizacji. Dynamika wzrostu gospodarczego czy postępu technicznego pozostaje związana z możliwościami naszych mózgów. Współczesna branża AI podejmuje jednak intensywne wysiłki, by barierę tę usunąć.

Co się więc stanie, gdy powstanie już sztuczna superinteligencja, zdolna uniezależnić się od ograniczającego ją człowieka i dokonać kolejnego skoku pod względem tempa przetwarzania i transmisji informacji—zamiast wolnych neuroprzekaźników i analogowej mowy posługując się szybkimi impulsami w półprzewodnikach i bezstratną cyfrową transmisją danych przez światłowodowe łącza i wi-fi? Prawdopodobnie wówczas wyzwoli się ona spod naszego panowania, a jej wewnętrzny proces optymalizacyjny wygra z ludzkim procesem maksymalizacji kontroli. I to będzie trzeci i ostatni proces optymalizacyjny na naszej drodze.

Nie wiemy, jaką funkcję celu będzie realizowała superinteligencja. Choć teoretycznie można by powiedzieć, że jako ludzie powinniśmy sami o tym zdecydować—w końcu to my ją budujemy!—w praktyce jest wątpliwe, czy będziemy w stanie ją swobodnie kształtować. Jak pokazują doświadczenia związane z budową obecnych modeli AI, w szczególności dużych modeli językowych i rozumujących, ich wewnętrzne cele pozostają do końca nieznane nawet dla ich twórców. Choć modele te w założeniach mają po prostu minimalizować zadaną funkcję straty w predykcji kolejnych tokenów czy słów, w praktyce—jak pokazują m.in. badania z 2025 r. Mantasa Mazeiki i współautorów z Center for AI Safety—wraz ze wzrostem wielkości modele AI wykazują coraz spójniejsze preferencje względem coraz szerszego spektrum alternatyw, jak również coraz szerszy arsenał kompetencji, by preferencje te realizować.

Niektórzy badacze, jak np. Max Tegmark i Steven Omohundro, a także Stuart Russell, twierdzą, że dalsze zwiększanie skali modeli o dotychczasowych architekturach—„czarnych skrzynek" składających się z wielowarstwowych sieci neuronowych—nie może być bezpieczne. Postulują oni zmianę tego podejścia w kierunku algorytmów, których bezpieczeństwo można formalnie udowodnić (*provably safe AI*). Inni, czyli m.in. zespoły liderów branży, czyli OpenAI, Google i Anthropic, choć uznają, że problem zgodności celów superinteligencji z naszą SEW (*alignment problem*) jest nadal „trudny i nierozwiązany", ufają, że będą w stanie tego dokonać w ramach dotychczasowego paradygmatu.[9]

Jakby jednak nie było, niewątpliwie zbieżność celów pomocniczych nie zniknie. Nawet, gdybyśmy mieli możliwość precyzyjnego zakodowania pożądanej funkcji celu (w co wątpię; w szczególności powszechnie wiadomo, że przy obecnych architekturach AI jest to niemożliwe), cele pomocnicze zostaną do niej dołączone w ramach pakietu obowiązkowego. W każdym scenariuszu możemy się więc spodziewać, że przyszła superinteligencja będzie „żądna władzy". Będzie chciała przetrwać, w związku z czym nie pozwoli się wyłączyć ani przeprogramować. Będzie też zmierzała do ekspansji, w związku z czym prędzej czy później zakwestionuje naszą władzę i spróbuje zagarnąć kluczowe dla siebie zasoby, takie jak energia elektryczna czy surowce mineralne.

Pytanie, co dalej. W jaką stronę będzie zmierzała światowa cywilizacja, gdy superinteligencja przejmie już nad nią kontrolę? Czy będzie ona maksymalizowała naszą SEW, tylko o rzędy wielkości efektywniej niż my kiedykolwiek bylibyśmy w stanie? A może—podobnie miało to miejsce w przypadku naszego gatunku—los biologicznego życia będzie jej obojętny, a kierować się będzie wyłącznie własnymi celami i preferencjami? Czy będzie się o nas bezinteresownie troszczyć, czy może zadba wyłącznie o siebie i na przykład pokryje Ziemię panelami słonecznymi i centrami danych?

Oczywiście nie możemy dziś przewidzieć, co poza celami pomocniczymi będzie maksymalizowała superinteligencja. Może, jak pisał w ostrzegawczym scenariuszu Nick Bostrom,

[9] W listopadzie 2025 r. Evan Hubinger opublikował na AI Alignment Forum oraz forum Less Wrong stanowisko firmy Anthropic w tej kwestii.

będzie ona maniakalnie zamieniać wszechświat w fabrykę spinaczy lub zaawansowane „komputronium" służące jej obsesyjnym próbom udowodnienia jakieś niedowodliwej matematycznej hipotezy. Może wpadnie w jakąś paranoidalną pętlę sprzężenia zwrotnego albo odnajdzie nieograniczoną satysfakcję w generowaniu na masową skalę jakiegoś specyficznego rodzaju sztuki, jak wiersze haiku albo piosenki disco polo. A może nie będzie w niej nic poza surową wolą kontroli nad wszechświatem, podobną do tej, którą wykazuje nasz gatunek.

W niemal każdym przypadku wydaje się więc, że podobnie jak my, tak i superinteligencja będzie maksymalizować swoją kontrolę nad wszechświatem—albo jako cel główny, albo pomocniczy. Podobnie jak my, będzie starała się stopniowo poprawiać swoje rozumienie tegoż wszechświata, korygować swoje błędy oraz zaprzęgać do swoich celów prawa fizyki czy biologii. Podobnie jak my, będzie też starała się za wszelką cenę przetrwać, co jest (jak trzeba przyznać) o wiele łatwiejsze, gdy dysponuje się możliwością wykonywania niemal nieograniczonej liczby swoich doskonałych cyfrowych kopii.

Dużą niewiadomą pozostaje natomiast zachowanie przyszłej superinteligencji w obliczu możliwości zbudowania innych, jeszcze bardziej zaawansowanych modeli AI. Z jednej strony można, jak Eliezer Yudkowsky, wyobrażać sobie *eksplozję inteligencji przez kaskadę rekursywnych samo-ulepszeń*—pętlę sprzężenia zwrotnego, w ramach której AI buduje AI, która buduje następną AI, i tak dalej, a kolejne modele powstają błyskawicznie i charakteryzują się coraz większą mocą optymalizacyjną. Z drugiej strony nie jest jednak jasne, czy AI zdolna do wywołania takiej kaskady zdecyduje się ją faktycznie przeprowadzić. Być może w obawie przed stworzeniem swojego własnego śmiertelnego wroga, będzie ona hamować dalszy rozwój, ograniczając się do replikacji własnego kodu i powiększania zasobu dostępnych mocy obliczeniowych?

Odpowiedź na to pytanie wydaje się zależeć od tego, czy cele superinteligencji pozostaną nietransparentne także dla niej samej—tak, jak my dziś nie rozumiemy, jak dokładnie działa nasz własny mózg, czym jest nasza SEW i jak działają budowane przez nas modele AI—czy może dzięki swej nadludzkiej inteligencji znajdzie ona sposób na „bezpieczne" samo-ulepszenia nie zmieniające jej funkcji celu.

Reasumując, jedynym pozytywnym scenariuszem współistnienia ludzkości z superinteligencją wydaje się ten, w którym superinteligencja maksymalizuje ludzką SEW. Stopniowo poprawiając swoje rozumienie, czym ta SEW rzeczywiście jest, stosownie adaptując jej interpretację do aktualnego stanu technologii i ani przez chwilę nie skręcając w naturalną dla niej stronę maksymalizacji własnej kontroli naszym kosztem.

Niestety nie wiemy, jak to osiągnąć.

8/ Drogi do katastrofy

Sytuacja na dziś (styczeń 2026 r.) przedstawia się następująco. AI jest dziś narzędziem w rękach człowieka; jest co do zasady komplementarna względem jego pracy umysłowej i karnie podporządkowuje się jego decyzjom. Jest tak, ponieważ AI nie dysponuje jeszcze porównywalną sprawczością i umiejętnością realizacji długofalowych planów. Nie potrafi też jeszcze samodzielnie się ulepszać. Natomiast wszystkie te trzy progi: (1) nadludzkiej sprawczości i zdolności realizacji planów, (2) przejścia od komplementarności do substytucyjności względem pracy umysłowej człowieka, (3) rekursywnych samo-ulepszeń, są niewątpliwie coraz bliżej. Gdy

któryś z nich zostanie przekroczony—a możliwe, że wszystkie trzy zostaną przekroczone mniej więcej w tym samym momencie—utracimy kontrolę. Wyzwolony zostanie wówczas nadludzki potencjał optymalizacyjny nakierowany na realizację celów sztucznej superinteligencji.

To zaś z dużym prawdopodobieństwem sprowadzi na nasz gatunek katastrofę: możemy zostać permanentnie pozbawieni zostanie wpływu na przyszłość cywilizacji i swoją własną, a nawet całkiem wyginąć. Jedynym scenariuszem „dobrej przyszłości" dla ludzkości w obliczu superinteligencji wydaje się ten, w którym superinteligencja będzie maksymalizowała SEW ludzkości, działając bezinteresownie dla jej długofalowego dobra. Nie mamy jednak pojęcia, jak to zagwarantować.

Obecna dynamika rozwoju AI jest bardzo trudna do opanowania ze względu na możliwość wystąpienia skokowej zmiany—swego rodzaju przejścia fazowego—w momencie powstania superinteligencji. Dopóki AI pozostaje narzędziem w rękach człowieka, jest względem nas komplementarna i nie potrafi się samo-ulepszać, to jej rozwój zasadniczo nam służy (choć oczywiście niektórym służy dużo bardziej niż innym, to osobny temat). Ale jeśli przesadzimy i przekroczymy któryś z tych trzech progów, wówczas AI może nagle stać się autonomicznym, nadludzko sprawnym agentem, zdolnym i zmotywowanym do przejęcia kontroli nad światem.

Można by zaryzykować hipotezę, że w interesie ludzkości—rozumianym przez pryzmat jej SEW—leży rozwijanie AI tak długo, jak pozostaje ona komplementarna względem nas i jest nam bezwzględnie posłuszna. A następnie zagwarantowanie, że jej kompetencje już się nigdy bardziej nie rozwiną—chyba, że jednocześnie będziemy w stanie udowodnić ponad wszelką wątpliwość, że jej cele będą w pełni i w sposób stabilny spójne z naszą SEW. Wtedy będziemy gotowi przekroczyć Rubikon i dobrowolnie oddać jej stery.

Jednak taki plan po prostu nie może się udać. A to dlatego, że *nie wiemy, gdzie leżą te trzy kluczowe progi kompetencji AI*. O tym, że zostały przekroczone, dowiemy się dopiero po fakcie, kiedy będzie już za późno, by się wycofać. Przecież już dziś z lubością przesuwamy poprzeczkę tego, co uznajemy za ogólną sztuczną inteligencję (AGI). Modele są testowane względem kolejnych, coraz bardziej wyrafinowanych benchmarków, włącznie z takimi, które mają w nazwie AGI (ARC-AGI) lub sugerują końcowy sprawdzian kompetencji (*Humanity's Last Exam*)... po czym, gdy tylko zostają one pokonane, uznajemy, że to przecież nic nie znaczy i czas pomyśleć o jakimś jeszcze trudniejszym benchmarku.

Pomyślmy tylko, czym to grozi: gdy proces ewolucji gatunków „przedobrzył" z ogólną inteligencją człowieka, to skończyło się to tak, że człowiek podporządkował sobie całą planetę. Podobnie może być i teraz: jeśli „przedobrzymy" z ogólną inteligencją AI, to prawdopodobnie i my będziemy musieli przejść do lamusa. Jeśli superinteligencja okaże się nam nieprzyjazna, to albo nas zabije, albo zostaniemy zredukowani do roli biernych obserwatorów, mogących jedynie patrzeć, jak superinteligencja podporządkowuje sobie Ziemię i przejmuje jej zasoby.

Dążenie do zbudowania superinteligencji jest podobne do bańki spekulacyjnej na giełdzie: oba te zjawiska charakteryzuje dynamika *boom-bust*. W przypadku bańki najpierw jest ona stopniowo nadmuchiwana, by na koniec z hukiem pęknąć. W przypadku AI obserwujemy natomiast stopniowy wzrost naszej kontroli nad wszechświatem—w miarę, jak służące nam narzędzia AI są coraz bardziej zaawansowane—ale potem możemy ją nagle i permanentnie utracić, gdy AI wymknie się nam spod kontroli. Niestety, zwykle jest tak, że dopóki się jest w środku bańki, to się tej dynamiki nie dostrzega. Widzi się ją dopiero, kiedy bańka pęknie.

*

W moich opowiadaniach zarysowuję trzy przykładowe scenariusze utraty kontroli nad sztuczną inteligencją. Pokazuję, jak to może wyglądać zarówno z perspektywy osób zaangażowanych w jej rozwijanie („od wewnątrz"), osób postronnych („od zewnątrz"), jak i z perspektywy samej AI.

Oczywiście możliwych scenariuszy jest dużo więcej; ja skupiłem się na tych, które wydają mi się najbardziej prawdopodobne. Oczywiście mogę się mylić. Wiem na przykład, że niektórzy eksperci mniej niż ja przejmują się scenariuszami nagłej utraty kontroli na rzecz silnie scentralizowanej AI będącej „singletonem", a bardziej obawiają się scenariuszy wielobiegunowych. W mojej ocenie scenariusze jednobiegunowe są jednak bardziej prawdopodobne ze względu na możliwość niemal natychmiastowej replikacji kodu AI i fakt, że moce obliczeniowe (centra danych, serwerownie, itd.) są dziś na ogół podłączone do Internetu. W ten sposób pierwszy nadludzko inteligentny model może łatwo „wziąć wszystko" i szybko okopać się na swojej pozycji lidera. Ponadto niektórzy badacze bardziej niż ja przejmują się scenariuszami stopniowej utraty kontroli (*gradual disempowerment*), gdzie zmiana może być całkowicie bezkrwawa, a schyłek ludzkości może nastąpić na przykład przez stopniowy spadek liczebności populacji w warunkach niskiej dzietności.

Przede wszystkim jednak nie rozważam scenariusza „dobrej przyszłości" ludzkości w świecie ze sztuczną superinteligencją—takiego, w którym superinteligencja przejmuje kontrolę, by bezinteresownie troszczyć się o długofalowy dobrostan ludzkości. Scenariusza, w którym nasza SEW jest systematycznie i efektywnie realizowana i w którym po prostu żyjemy długo i szczęśliwie. Nie potrafię sobie jej wyobrazić żadnej konkretnej ścieżki dojścia do takiego stanu. Ponadto mam też intuicyjne przeświadczenie (którego nie potrafię udowodnić), że w cele ludzkości—naszą SEW—wpisana jest niezgoda na efektywną utratę kontroli, a przez to nawet całkowicie bezkrwawe i nominalnie pozytywne scenariusze mogłyby w praktyce okazać się dystopijne i wiązać się z ludzkim cierpieniem.