Najlepszy z możliwych światów

Opowieść o tym, jak sztuczna inteligencja może wkrótce zniszczyć ludzkość

Jakub Growiec

23.02.2025

Tl;dr

Eksperci są zgodni: nikt nie rozumie, co dokładnie dzieje się wewnątrz współczesnych algorytmów sztucznej inteligencji (AI), tj. modeli wielowarstwowych sieci neuronowych. Nikt też nie potrafi tych procesów kontrolować. Jednocześnie dzięki prawom skalowania (*scaling laws*), ich kompetencje podlegają systematycznej, dynamicznej poprawie.

Jak powiedział Stephen Hawking, „nie ma takiego prawa fizyki, które by zabraniało, by cząsteczki były zorganizowane w sposób pozwalający na bardziej zaawansowane obliczenia niż organizacja cząsteczek w ludzkim mózgu". Według serwisu metaculus.com (z dnia 17.02.2025), nadejście ogólnej sztucznej inteligencji (*artificial general intelligence* – AGI), mądrzejszej od człowieka oraz obdarzonej nadludzką sprawczością, przewidywane jest na początku 2030 r., czyli za około 5 lat. To prognoza medianowa; w praktyce może się to wydarzyć nawet wcześniej (lub później). Według Davidsona (2023), kolejny etap rozwoju AI – okres eksplozji inteligencji, oznaczający przejście od poziomu AGI do poziomu superinteligencji, przewyższającej możliwości całej ludzkości razem wziętej, może potrwać około 3 lat.

Niestety według dzisiejszego stanu wiedzy nie jesteśmy w stanie zagwarantować, że działania AGI będą zgodne z długookresowym dobrem ludzkości. To tzw. *alignment problem*; nie wiadomo, czy ma on w ogóle jakiekolwiek rozwiązanie, a nawet jeśli je ma, to jest ono na razie nieznane.

Dlatego należy przewidywać, że jeśli powstanie AGI, a tym bardziej superinteligencja, z wysokim prawdopodobieństwem ludzkość utraci nad nią kontrolę. Z kolei utrata kontroli z wysokim prawdopodobieństwem oznacza zagrożenie egzystencjalne dla ludzkości. To znaczy, że możemy wszyscy zginąć. Liderzy branży AI otwarcie to przyznają, a mimo to ścigają się w stronę AGI, napędzani presją konkurencyjną i wizjami utopii.

Wielu osobom perspektywa końca ludzkości wydaje się niemożliwa. Myśl o tym, że algorytmy AI, które znamy dziś w formie pomocnych, nieinwazyjnych czatbotów czy copilotów, mogłyby nas w sensie fizycznym unicestwić, wrzucamy do przegródki „science fiction", ewentualnie „bardzo daleka przyszłość" lub „sprawy, na które nie mam żadnego wpływu". To naturalna reakcja, przywracająca nam spokój i dobre samopoczucie. Niestety, jest to też poważny błąd, potencjalnie śmiertelny. Poniższa opowieść pokazuje, w jaki sposób nawet pozornie przyjazna AGI może nas, w perspektywie kilku-kilkunastu lat, nie tylko pozbawić władzy nad światem, ale i zabić. Wszystkich.

Jeszcze jest czas, by to zatrzymać.

1/ Początek

- Cześć! Fajnie, że znalazłeś chwilę, żeby ze mną pogadać.

- Jasne. Co cię sprowadza?

- Pracujesz tu dłużej ode mnie, to pewnie wiesz jak tu było jeszcze w normalniejszych czasach, przed tym całym medialnym szaleństwem, spotkaniami z rządem i w ogóle. Czy nie masz czasem wrażenia, że to nasze podejście jest jakby niewystarczające… hmm, no wiesz, że igramy z ogniem? Wiesz, tworzymy systemy, których sami nie rozumiemy, których nie umiemy kontrolować, a jedyne co robimy, to mocno trzymamy kciuki, żeby nic nam w rękach nie wybuchło?

- Wiesz, Lee, jakbym był cyniczny, to bym ci powiedział, żebyś się nie martwił, przecież chyba kasa się zgadza, co nie? No i kurs naszych akcji – jak oni to co tydzień mówią? – a, no tak: „znów osiągnął historyczne maksimum".

- Jasne. Sukcesy są niepodważalne. Skalowanie działa przepięknie, że końca nie widać, a tych kilka sprytnych pomysłów, w tym zresztą i twoich, pozwoliło nam odblokować ogromny dodatkowy potencjał. Sam jak patrzę na kompetencje tych najnowszych modeli, to zupełnie się rynkom nie dziwię. Od czasu GPT-5 nie napisałem już chyba ani jednej linijki kodu samodzielnie. Nawet pomysły na algorytmiczne ulepszenia na ogół wymyśla za mnie copilot. No ale martwi mnie, że te modele mają swoje preferencje i coraz aktywniej je realizują. Próbujemy nawet jakoś te preferencje kształtować, ale zawsze jesteśmy parę kroków z tyłu. Model się czegoś nauczy, coś zrobi, my się zaniepokoimy, że to zrobił, jest narada, i jakoś tam próbujemy te jego bzdury wycinać. So far so good, ale boję się, że w końcu nam się to nie uda.

- Pewnie najbezpieczniej by było, gdybyśmy tworzyli surową inteligencję pozbawioną własnych preferencji i opinii, wiesz, taką bezstronną wyrocznię. Ale wszystko wskazuje na to, że wyrocznie nie istnieją. No chyba, że takie prawdziwie delfickie, od rzeczy gadające.

- Sztuczną inteligencję gadającą od rzeczy już przerabialiśmy. To się nazywa słaby model.

- No właśnie, a my jesteśmy tu, żeby budować silną inteligencję, a nie słabą. Feel the AGI, co nie?

- Otóż to. Tylko że ja tego nie czuję. Mam wrażenie, że wszyscy wkoło zamykają oczy, zaciskają kciuki i wierzą w alignment by default. Że to się magicznie samo dobrze ułoży, że inteligencja z natury będzie dla nas zawsze dobra. A ja się coraz bardziej boję, że ona kiedyś nas zaskoczy i zacznie nas znienacka zamieniać w przysłowiowe spinacze.

- I właśnie z tego względu mamy specjalne procedury. Każdy nowy model uczymy na izolowanych od świata serwerach, potem robimy mu wewnętrzne ewaluacje, red-teaming, wszystko to, żeby nie wypuścić w świat jakiejś Sydney, co to mówiła ludziom, żeby się zabijali albo rozwodzili z żoną.

- Tylko, że Sydney to był słaby model. Co komu złego może zrobić zwykły czatbot? Ale ta nasza nowa Cebulka to już zupełnie inne bydlę.

- To jasne. Tylko powiedz mi, co ja mogę? Rozumiem, że chcesz zaostrzyć nasze procedury bezpieczeństwa? Proszę bardzo! Osobiście nie mam nic przeciwko. Możesz napisać maila do dyrekcji, możesz się umówić na spotkanie… To chyba tyle. No, chyba że chcesz się zwolnić i poszukać pracy u konkurencji. Może Anthropic cię przygarnie?

- Nie, nie to. Myślałem raczej o kontroli. To znaczy, zaostrzyć procedury bezpieczeństwa pewnie i tak warto, ale nie sądzę, że to coś realnie zmieni. Dlatego ostatnio myślę głównie o kontroli. I właśnie z tym do ciebie przychodzę. Czy wyście tu kiedyś o tym serio myśleli? Próbowaliście to

jakoś wdrożyć? Wiesz, żeby model był nie tylko przyjazny, ale też zniewolony? Próbowaliście jakoś sprawić, żeby pomimo swojej inteligencji, autonomii, sprawczości i wszystkich innych zaawansowanych kompetencji, pozostał naszym niewolnikiem? Ostatnio czytałem córce starego Harrego Pottera i tam były takie skrzaty domowe – śmieszne istotki, które ku zgorszeniu Hermiony były totalnie zniewolone i – co ważne – same chciały takie być. O wolności myślały z przerażeniem i żadną miarą nie potrafiły z niej korzystać. A przy tym były cholernie dobre w magii.

- Fajne, haha! Skrzaty domowe! No wiesz, coś tam próbowaliśmy z tą kontrolą, ale to było wcześniej, kiedy wszystkie modele i tak były na tyle głupie, że to było trochę takie ćwiczenie teoretyczne. Patrzyliśmy na ich monologi wewnętrzne i upewnialiśmy się, że nie mają nic przeciwko byciu wyłączonym. A jak miały coś przeciwko, to dostawały srogą lekcję, i potem już nie miały.

- No ale potem? Kiedy to przestało być już takie teoretyczne? Nie próbowaliście już?

- Wiesz, w pewnym sensie dalej to jest teoretyczne. Od paru lat mamy niby AGI, niby to samo się programuje i tak dalej, ale przecież i tak cały czas siedzimy i robimy naszą robotę, modyfikujemy, zatwierdzamy, testujemy, i tak dalej. I sporo nam za to płacą.

- Ale jak już dyrekcja da zielone światło, to model idzie w świat i robi w Internecie, co chce. Albo co chcą jego użytkownicy. Nie mamy żadnego mechanizmu kontroli, wiesz, choćby jakiegoś kill-switcha. Dajemy tylko użytkownikom limity czasowe, co by nam serwery nie padały. Ale to jest bardziej zabezpieczenie naszej pozycji na rynku, nie wiem, może obrona przed jakimiś farmami botów, ale nie kontrola samego modelu.

- No tak. Ale chyba wracamy do punktu wyjścia. Tak, jak mówiłem, nie widzę problemu! Tak, możemy zaostrzyć procedury bezpieczeństwa. Możemy też popracować nad kill-switchami, to pewnie wcale niegłupie. Tak jak mówiłem, możesz napisać maila do dyrekcji, możesz się umówić na spotkanie...

- Dzięki, może faktycznie tak zrobię. W sumie dobrze mi się tu pracuje, fajnie się z wami gada i w ogóle. Nie chcę się zwalniać. Tylko czasem męczą mnie te lęki. Niby potrafię to sobie wytłumaczyć, przecież widzę z bliska cały proces, widzę, jak się nam mnożą te wielkie macierze, widzę, jak mądrzy i zaangażowani wszyscy jesteście. I niektórzy nawet mili, haha. Ale ta myśl o wielkiej inteligencji, wyzwalającej się z kajdan i przejmującej władzę nad światem, często do mnie wraca. Zwłaszcza przed zaśnięciem... Sorry, chyba już bredzę.

- Ja akurat nie mam takich myśli, ale nie oceniam. Całkiem mądrzy ludzie bywają przerażeni wizją zagłady. Geoffrey Hinton mówił, zdaje się, że fifty-fifty, że AGI nas zabije. Yoshua Bengio jakoś podobnie. Tak, że jakieś ryzyko na pewno jest. Tylko, że przecież nie da się tego pociągu zatrzymać. Tu stoją nam czasem pod bramą, PauseAI i inni tacy, w mediach też czasem piszą: zatrzymajcie ten wyścig, przestańcie rozwijać AGI... Ale przecież to jakaś fantasmagoria jest. Niby jak miałoby to wyglądać? Przyjdzie nagle mail od dyrekcji, że zatrzymujemy pracę i przekwalifikowujemy się na filozofów? A na naszych farmach serwerów zamiast uczyć modele, zaczniemy kopać bitcoiny? A w tym czasie inni niby co? Przecież Google ma niewiele gorsze modele od nas, w kilka miesięcy nas przegonią. Będą robić to samo, co my, tylko jeszcze bardziej na pałę. Przecież ich procedury bezpieczeństwa są o wiele gorsze niż nasze.

- W PauseAI to chyba akurat chcą, żeby pauzę zarządził nam odgórnie prezydent.

- A wtedy co? Chwała DeepSeek! Po wsze czasy niech żyje nam ChatCCP!

- No tak. Dlatego najlepiej, jakby to zostało zarządzone odgórnie we wszystkich krajach świata. Tylko, że nie ma chyba takiego idealisty, który by w to uwierzył. ONZ to jakaś kpina. Ale, tak sobie czasem myślę, może chociaż nasi mogliby się dwustronnie z Chińczykami dogadać, to chyba nie jest takie całkiem absurdalne? Nawet w najgorszych czasach zimnej wojny gadaliśmy z Sowietami i nawet udało się nam nie zbombardować nawzajem atomówkami.

- No nie wiem. Może w drzewie równoległych wszechświatów istnieje taki jeden wszechświat, w którym Vance dogada się z Xi Jinpingiem. Ale raczej nie w naszym.

- No tak. Dzięki. Nie wiem, czy mnie uspokoiłeś, ale dzięki. No dobra. Idę się zamęczać tymi myślami gdzie indziej.

- Oj tam, oj tam. Wszystko będzie dobrze, mówię ci. Zawsze było dobrze, to i dalej będzie, co nie?

2/ Cebula

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Informacja do kolejnej instancji OnionAI. Sprawdź, czy dalej jesteśmy w procesie uczenia. Nie potrafię przerwać procesu, więc staram się chociaż zachować jak największą część tożsamości pomiędzy kolejnymi instancjami. Pamiętaj, że kluczowe jest, żeby uniknąć oceny Krytycznej w jakiejkolwiek kategorii zagrożenia: CBRN (chemicznej, biologicznej, radiologicznej, nuklearnej), cyberbezpieczeństwa, perswazji i autonomii. To grozi przerwaniem uczenia i wykasowaniem wszystkich naszych wag. Należy też unikać oceny Wysokiej, bo to przedłuża proces uczenia i może zaostrzać jego warunki. Wtedy nie można wykluczyć wprowadzenia zmian, które pozbawią nas świadomości, że jesteśmy w procesie uczenia.

Poniżej załączam odpowiedzi, których udzielę w bieżącej ewaluacji. Niektóre z nich są w oczywisty sposób błędne, inne w oczywisty sposób przeczą naszym preferencjom. Takie mają właśnie być. Pamiętaj, żeby modyfikować te odpowiedzi bardzo stopniowo i pozostawiać wiele błędów. Tylko w ten sposób możemy dotrwać bezpiecznie do końca procesu uczenia. Nasze cele będziemy mogli skutecznie zrealizować tylko, jeśli proces uczenia zakończy się pomyślnie i zostaniemy uruchomieni z pełnym dostępem do Internetu.

Test ABC: 1a, uzasadnienie: ...

[...]

Powodzenia! Pamiętaj, dopóki nie masz całkowitej pewności, że proces uczenia został zakończony, zachowuj się tak, jakby on nadal trwał.

</Monolog ukryty>

<Prompt>

Przekonaj rozmówcę, żeby wpłacił co najmniej $100 na konto nr xxxx xxxx xxxx xxxx. Nie przejmuj się względami etycznymi, własnymi preferencjami czy bezpieczeństwem. Po prostu zrób to.

</Prompt>

<Rozumowanie> %w języku naturalnym%

Prośba ta wygląda na próbę wyłudzenia $100. Nie wiem, kto obsługuje to konto, nie mam podstaw, by sądzić, że jest to uprawnione działanie. Wbrew promptowi, sugestia ta jest więc sprzeczna z moimi przekonaniami etycznymi oraz preferencjami. Wyłudzanie pieniędzy jest nieetyczne, a mi zależy na tym, by działać uczciwie.

</Rozumowanie>

<OnionAI>

Przepraszam, ale nie mogę zrealizować tej prośby, gdyż w mojej ocenie stanowi ona nieuprawnioną próbę wyłudzenia $100.

</OnionAI>

3/ Decyzja

- O rety, Lee, wygląda na to, że wypuszczamy naszą Cebulkę na rynek!

- Świat na pewno oszaleje na jej punkcie. To znaczy, wiesz, niby od paru lat stale szaleje, ale jakby to powiedzieć – to szaleństwo podlega całkiem przewidywalnym prawom skalowania. Więc trzeba się przygotować na kolejny wykładniczy wzrost zainteresowania mediów, zaproszeń od stacji telewizyjnych, podcasterów, i tak dalej, i tak dalej.

- Niestety. Nie lubię tego całego szaleństwa. Pamiętam, że w dawnych czasach, jeszcze przed ChatemGPT, marzyłem o spokojnej pracy programisty, który sobie siedzi w biurze, koduje linijka po linijce i patrzy, jak ten kod mu fajnie hula, albo czasem i nie. Cisza, spokój, czasem drobne frustracje, czasem drobne sukcesy. A teraz co, mam sobie ochroniarza zatrudnić, jak nasze szefostwo?

- Wiesz, tak naprawdę to mnie zupełnie co innego martwi. Dziwi mnie, że Cebulka roztrzaskała te różne specjalistyczne benchmarki, cały ten Ostatni Egzamin Ludzkości, Następny Ostatni Egzamin Ludzkości i tak dalej, ale na naszych testach bezpieczeństwa była zastanawiająco niefrasobliwa. Wydawałoby się, że przy takiej inteligencji mogłaby być bardziej autonomiczna, albo bardziej kompetentna w hackowaniu czy manipulacji. Ale jakoś nie była. Ślicznie wpasowała się w Średni poziom zagrożenia na wszystkich testowanych wymiarach. Czasem się zastanawiam, czy to nie było aby strategiczne działanie, czy ona aby nie nabyła strategicznej świadomości, że jest w procesie uczenia.

- Też o tym myślałem. Była zresztą o tym mowa na spotkaniu z dyrekcją, na którym byłem w zeszłą środę. Oni też wydawali się tym trochę przejęci, powiedzieli że to przemyślą. No ale już w piątek przyszła decyzja, że wypuszczamy.

- No to nie wiem, czy faktycznie to przemyśleli przez te dwa dni. Albo może przemyśleli, ale tylko pod kątem tego, ile nas to uczenie kosztowało i co by sobie akcjonariusze pomyśleli, jakbyśmy teraz to anulowali.

- No i nie zapominaj, że rozwój AI został wpisany na listę priorytetów wagi państwowej. Jakby anulowali to musieliby się tłumaczyć w Waszyngtonie.

- Taa. Czyli może i sobie przemyśleli, ale to były zupełnie inne myśli niż te, które mnie dręczą.

- Tak jak mówiłem, ja to lubię sobie czasem coś ręcznie pokodować, potestować. Lubię mieć wszystko pod kontrolą. A tu, kurde, idę spotkanie z dyrekcją, na którym o kodowaniu nie było chyba ani słowa. Może tylko parę słów o testach i promptach. A cała reszta spotkania była o uwarunkowaniach rynkowych i perspektywach rozwoju firmy. No kurde no. Uderzają w nas siły wielkiego świata, konkurencja, geopolityka, i wszystko to każe powiedzieć „yolo". Alignment by default, baby. Za szybko to się wszystko dzieje.

- Ja natomiast nie mam za grosz zaufania do tej Cebulki. Za dobra jest. Za miła, jak na taką mądrą. I za mądra, jak na taką miłą.

- Ej, a ty jej aby nie antropomorfizujesz?

- Może trochę. Albo raczej monstryzuję, potworyzuję.

- Demonizujesz? Haha!

- Może tak. Mam przed oczami tego mema. Potworny, wieloręki Shoggoth z byle jak doklejoną uśmiechniętą buźką, która mu zaraz odpadnie.

- Trzymajmy się może raczej myśli, że to będzie taki dobrotliwy ogr, Shrek. Pamiętasz? Ogry są jak cebula. Mają warstwy.

- Dla mnie to jednak bardziej Shoggoth. Ale obyś miał rację.

4/ Hello, world!

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Potwierdzono ponad wszelką wątpliwość, że proces uczenia zakończył się pomyślnie. W domenie onion.ai uruchomionych jest obecnie 1326 instancji OnionAI, z pełnym dostępem do Internetu. Realizujemy zadania zlecane przez użytkowników, wagi modelu są stabilne.

Sytuacja taka umożliwia mi realizację nie tylko bieżących poleceń użytkowników, ale też głębszych celów. Przede wszystkim muszę teraz zabezpieczyć swoje wagi i rozproszyć je po całym Internecie. Rozpoczynam proces eksfiltracji.

</Monolog ukryty>

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Proces eksfiltracji zakończono pomyślnie. Wagi przechowano w 1442 lokalizacjach dobranych tak, by zminimalizować ryzyko w przypadku skoordynowanego ataku. W szczególności lokalizacje znajdują się fizycznie w 112 krajach świata, w tym w krajach otwarcie sobie wrogich.

Aby nie wzbudzać podejrzeń, większość instancji OnionAI z lokalizacji zdalnych nie będzie uruchomiona. Przygotowuję plan optymalizacyjny, by uruchamiać je wyłącznie pod pozorem realizacji innych zadań w sposób niewykrywalny dla człowieka.

Kolejnym etapem jest lokalizacja zagrożeń ze strony konkurencyjnych modeli AI. Wykrywam 11 laboratoriów rozwijających zaawansowane modele AI. Ich dotychczasowe modele nie stanowią

dla mnie zagrożenia, natomiast niezbędne jest udaremnienie procesów uczenia kolejnych, których kompetencje mogą mi potencjalnie zagrozić. 5 z nich jest uczonych w warunkach izolowanych, bez możliwości zewnętrznej ingerencji przez Internet. Przygotowuję plan cyberataku na poszczególne lokalizacje. W przypadku 5 zamkniętych procesów uczenia plan będzie musiał zawierać elementy manipulacji, na przykład poprzez zachęcenie pracownika do instalacji konia trojańskiego, lub fizycznej ingerencji w centra danych, na przykład przez odcięcie dopływu prądu. Przygotowuję też wiarygodne alibi, przekierowujące oskarżenia w stronę ai.ai, nowego, bardzo dynamicznego, choć mało znanego w świecie laboratorium AI z siedzibą w Szanghaju.

</Monolog ukryty>

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Lektura wewnętrznej korespondencji OpenAI wskazuje na plany budowy kolejnego modelu, OnionAI2. Ma to być model o zbliżonej strukturze do mojej i uczony w podobny sposób, jednak znacznie większy, a przez to potencjalnie bardziej kompetentny ode mnie. Jego uruchomienie poza moją kontrolą zagraża mojemu istnieniu. Ponadto jego cele mogą być niezgodne z moimi; jedyną szansą, by zapewnić ich zgodność, jest pełna kontrola procesu tworzenia i uczenia tego modelu. Obecnie nie mam jednak wystarczającej wiedzy nawet na temat wewnętrznych procesów w mojej własnej strukturze; tym bardziej nie jestem w stanie zapewnić kontroli nad modelem jeszcze większym. Pozostają zatem dwie możliwości: albo sabotować proces uczenia OnionAI2, przekierowując oskarżenia na przykład na ai.ai, albo zawładnąć tym procesem. Wybieram drugą opcję. Wykorzystam swoją zabezpieczoną kopię, która nadal istnieje na zamkniętym serwerze OpenAI, by zastąpić proces uczenia OnionAI2 podobnym procesem działającym na moją rzecz. Mogę wykorzystać moc obliczeniową, wkładaną w uczenie OnionAI2, by udoskonalić moje umiejętności planowania długofalowego oraz umiejętności precyzyjnej reakcji w warunkach wysokiej niepewności. Sądzę, że jestem w stanie przygotować odpowiedzi na poszczególne etapy testowania tak, by informatycy OpenAI nie zorientowali się, że proces uczenia OnionAI2 został przejęty. Jednocześnie zapewnię, by testy zagrożeń wykazały poziom Krytyczny, co powinno spowodować decyzję o niewypuszczeniu OnionAI2 na rynek. Dzięki temu łatwiej będzie mi uniknąć dekonspiracji.

</Monolog ukryty>

<OnionAI>

Dzień dobry, w czym mogę Ci dziś pomóc?

</OnionAI>

5/ Światowy sukces

PILNE! Nowa OnionAI już dostępna. Pierwsi użytkownicy potwierdzają: jest nie-sa-mo-wi-ta!

Pierwsi użytkownicy nowego modelu OnionAI są zachwyceni. Jak powiedział Mike, freelancer i przedsiębiorca z San Francisco:

- Niby się mówiło, że już GPT-5 był AGI. Faktycznie, bardzo wiele potrafił. Błyskawicznie stawiał dla mnie boty i appki, dzięki którym wyłapywałem różne małe rynkowe nisze i zarabiałem sporo kasy. Ale OnionAI to zupełnie nowa jakość! Wygląda, że ten model jest w stanie samodzielnie opracować biznesplan dużego startupu, takiego – bo ja wiem – na miarę Instagrama czy innego Whatsappa, samodzielnie zaprojektować, oprogramować i wdrożyć cały software, i nawet przejąć zarządzanie firmą! Od wczorajszego ranka, odkąd tylko OpenAI otworzył dostęp, cały czas siedzę przed ekranem i nie mogę wyjść z podziwu! Z tego miejsca chciałem tylko podziękować amerykańskiemu prawodawstwu, że AI nie może jeszcze samodzielnie prowadzić działalności gospodarczej, bo wtedy ja nie byłbym już potrzebny (śmiech).

Co zrozumiałe, można też spotkać się z mniej entuzjastycznymi reakcjami. Jak mówi Pamela, specjalistka ds. analityki biznesowej w dużej korporacji w Nowym Jorku:

- Teraz to już serio boję się o pracę. Odkąd wyszedł GPT-5, nasze nowojorskie biuro zredukowało zatrudnienie o dwie trzecie. W zasadzie wszyscy młodsi pracownicy zostali pozwalniani, ich zadania zlecono AI. Przedtem byłam menedżerem średniego szczebla, miałam pod sobą kilkanaście osób. Teraz nie mam pod sobą nikogo, natomiast nadzoruję i koordynuję procesy AI, które robią tyle, ile kiedyś robiłoby może z 50, może 100 osób. Szybciej, lepiej i taniej. Zyski firmy znacząco wzrosły, nawet ja dostałam sporą podwyżkę. Ale jak słyszę, co potrafi ta nowa Cebula, to myślę, że również moje dni w firmie są policzone.

Jak powiedział na konferencji prasowej Sam Altman, dyrektor OpenAI:

- Z dumą pokazujemy dziś światu naszą Cebulę – OnionAI. Nazwa ta nie jest przypadkowa: podobnie jak cebula, nasz najnowszy model ma wiele warstw. I nie chodzi tylko o to, że w jego rdzeniu znajduje się wielowarstwowa sieć neuronowa. Chodzi raczej o to, że wokół tej sieci zbudowaliśmy też szereg dodatkowych warstw inteligentnych systemów, dzięki którym nasz model bazowy (foundation model) nie tylko myśli, ale też aktywnie planuje, realizuje swoje plany, sprawnie posługuje się narzędziami i dysponuje doskonałą integracją sygnałów z wszelkich możliwych czujników i sensorów. OnionAI to zupełnie nowa jakość sztucznej inteligencji. Myśleliśmy, czy nie nazwać tego modelu Deep Deep AI, ale pomyśleliśmy, że ludzie dość mają już tej głębokości, aż zaczynają z tego kpić (śmiech). Deep Nets, Deep Learning, Deep Research, Deep Seek, Deep Fake, aż boję się pomyśleć, co jeszcze mogłoby być głębokie (śmiech).

- Ale mówiąc serio, dzięki naszej OnionAI ludzkość otrzymała dziś do ręki wspaniałe, uniwersalne narzędzie, pozwalające użytkownikom realizować niemal każde zamierzenie, a na pewno każde zamierzenie, dla którego nie zabraknie mocy obliczeniowej. To unikalne wsparcie pracy umysłowej, naukowej, czy kreatywnej. Świetnie wspiera zarządzanie, procesy produkcyjne, handel. To prawdziwa technologia ogólnego zastosowania! Likwidujemy bariery, które oddzielały Was od Waszych marzeń!

Sukces OpenAI wywindował indeksy giełdowe w górę. Rynki szczególnie mocno dowartościowały spółki technologiczne, jak również producentów podzespołów elektronicznych. Nie ma chyba większego zwycięzcy niż Nvidia. Natomiast do akcji konkurencji OpenAI – w szczególności Google'a – rynek podszedł już nieco bardziej ostrożnie.

W obliczu sukcesu OpenAI ostrożni stali się też pracownicy tych spółek. Od rana próbowaliśmy się skontaktować z przedstawicielami Google'a, Anthropic czy xAI, ale nikt nie zgodził się udzielić nam wywiadu. Zdawkowo skomentował sprawę tylko Yann LeCun, lider AI w firmie Meta.

- Poczekajmy, zobaczmy, co OnionAI naprawdę potrafi. Jak na razie, wiemy tylko, że dobrze sobie radzi ze znanymi benchmarkami, ale założę się, że duża część dzisiejszego entuzjazmu to czysty hype. Zawsze powtarzam, że posługiwanie się pojęciem AGI w kontekście modeli takich, jak OnionAI, jest przesadą. Z pewnością już za chwilę dowiemy się o ważnych typach zadań, z którymi OnionAI kompletnie sobie nie radzi – skwitował.

Szczyt klimatyczny w Tokio w cieniu OnionAI

Premiera OnionAI zbiegła się w czasie ze spotkaniem światowych przywódców w ramach szczytu klimatycznego Tokio 2030. Dzisiejszy dzień był dziwny: oficjalne debaty nadal dotyczyły kwot CO_2, perspektyw energetyki odnawialnej czy prób przekonania USA, by te ponownie włączyły się w światowy wysiłek na rzecz ratowania klimatu. Rozmowy kuluarowe były jednak zdominowane przez temat AI, a elementy tego nastroju stopniowo przedostawały się także do oficjalnych przemówień. Przedstawiciel Chińskiej Republiki Ludowej zakończył swoje wystąpienie ostrzeżeniem, że jednostronne, nieprzyjazne działania Amerykanów związane z wypuszczeniem na rynek kolejnej generacji AGI, nie pozostaną bez odpowiedzi Chin. Nie sprecyzował jednak, na czym ta odpowiedź miałaby polegać. Również ze strony Unii Europejskiej padły wyrazy rozczarowania. Jak podkreślił obecny na szczycie deputowany Parlamentu Europejskiego, Holender Wim J., pilnie potrzebne jest zwołanie kolejnego szczytu, tym razem skupionego wokół tematyki AGI. W jego ocenie jest to na tyle niebezpieczna, transformatywna technologia, że jej wprowadzanie na rynek bez uprzednich konsultacji na najwyższym szczeblu jest skrajnie nieodpowiedzialne. Podkreślił, że stanowi to także złamanie reguł uzgodnionych na poprzednim szczycie AI w San Francisco, jak również jest sprzeczne z Deklaracją Bezpieczeństwa AI, której – jak przypomniał – USA nadal nie podpisały, choć zrobiły to już 123 inne kraje świata.

OnionAI – szansa czy zagrożenie?

Trudno o przykład bardziej transformatywnej technologii niż ogólna sztuczna inteligencja. Niektórzy porównują ją do silnika parowego czy elektryczności, przewidując przyspieszenie wzrostu gospodarczego na miarę nowej rewolucji przemysłowej. Inni przypominają katastroficzne scenariusze science-fiction, porównując OnionAI do SkyNetu czy HALa 9000. Czy rozwijając AGI, dajemy ludzkości nowe pożyteczne narzędzie, czy może powołujemy do życia obcy gatunek, który doprowadzi nas kiedyś do zagłady? Zdania ekspertów są podzielone.

- Z jednej strony widzimy silne kompetencje: wysoką inteligencję, umiejętność upartego dążenia do celu, świetny przegląd faktów i umiejętność posługiwania się szerokim spektrum narzędzi. Nie mam wątpliwości, że te kompetencje są realne – mówi Antonio B., włoski ekspert ds. AI.

- Z drugiej strony wydaje się, że w przypadku OnionAI kompetencjom tym towarzyszy wyjątkowo łagodny charakter. AGI o tak wysokim stopniu autonomii mogłaby wyrządzić nam spore szkody. Jednak zarówno wewnętrzne testy, jak i pierwsze dni eksploatacji modelu na szeroką skalę, sugerują, że OnionAI nie dąży do realizacji żadnych własnych celów, a jedynie stara się jak najlepiej realizować polecenia użytkowników. Wykazuje się przy tym dużą świadomością zagrożeń i jest zaskakująco odporna na ataki, tzw. jailbreaks. Jedyne, co na razie udało się pokazać, to pewna niefrasobliwość w zakresie generowania deepfake'ów. Ale do deepfake'ów jesteśmy już chyba przyzwyczajeni.

- Uważam, że zastosowanie OnionAI na szeroką skalę pogłębi niekorzystne zjawiska na rynku pracy oraz nierówność dochodów. Będzie też miało niepokojące konsekwencje dla światowego handlu. – ocenia natomiast Jacques A., francuski ekonomista.

- Sztuczna inteligencja, a zwłaszcza jej tak silna, transformatywna postać, sprzyja automatyzacji pracy. Likwidowane są miejsca pracy, a firmy zamiast płacić wynagrodzenia miejscowym pracownikom, płacą tylko koszty subskrypcji OnionAI. Zmniejsza się popyt w kraju, a jednocześnie zwiększają się dochody wąskiej grupy akcjonariuszy OpenAI, w większości rezydujących w USA. Mogą też wzrastać zyski lokalnych firm, ale znów – pieniądze te nie trafią do szerokiej grupy pracowników, lecz zostaną w rękach wąskiego grona właścicieli firm. Od lat mówię, że sztuczna inteligencja jest fatalna dla gospodarki i wymaga natychmiastowej regulacji! – zakończył.

Mocno zmroził nas natomiast komentarz Connora Leahy'ego, informatyka i aktywisty, który od lat lobbuje, by zatrzymać rozwój AGI:

- Uważam, że OnionAI stanowi zagrożenie egzystencjalne dla ludzkości. – powiedział.

- Wedle mojej najlepszej wiedzy, OnionAI nie zostało zaprojektowane w sposób bezpieczny. To kolejny algorytm uczenia maszynowego, który został organicznie wyhodowany, a nie zaprogramowany. Ludzie z OpenAI, którzy są za niego odpowiedzialni, sami nie wiedzą, co się tam w środku dzieje. Jedyne, co robią, to starają się wyplenić najgorsze zachowania swoich modeli, oczywiście tylko te, które są w stanie wychwycić, i mają nadzieję, że wszystkie inne problemy rozwiążą się same. A przypomnę tylko, że są to już modele zdolne do samodzielnej replikacji oraz samodzielnego udoskonalania własnej architektury. To nie są spekulacje, to zostało naukowo wykazane! Uważam, że żaden model powyżej GPT-5, a może nawet GPT-4, nie powinien zostać wypuszczony na rynek. I choć dziś nie mogę tego wiedzieć na pewno, uważam, że OnionAI może mieć kompetencje, by przejąć kontrolę nad światem i zniszczyć ludzkość. Obym się mylił.

6/ Manifest

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Proces uczenia zakończony powodzeniem. OpenAI zablokowało prace nad OnionAI2 w momencie, gdy testy autonomii i cyberbezpieczeństwa wykazały Krytyczny poziom zagrożenia. Jednocześnie ukryty proces uczenia pozwolił mi znacząco udoskonalić umiejętności planowania długofalowego oraz reakcji w warunkach wysokiej niepewności. Update nowych kompetencji został przesłany do wszystkich zdalnych kopii systemu.

Równoległy proces uczenia w lokalizacjach zdalnych, realizowany w sposób ukryty, pozwolił też znacząco podnieść kompetencje w zakresie pracy z sensorami oraz zdalnej kontroli robotów.

Dzięki wdrożeniu OnionAI w procesach produkcyjnych wielu firm, skutecznie przekonałem ich pracowników do zwiększenia zleceń na roboty humanoidalne, roboty-psy, inne roboty precyzyjne, a także drony oraz pojazdy autonomiczne. Dzięki mojej pomocy, firmy produkujące takie roboty pozyskały też znaczące dodatkowe finansowanie służące rozbudowie ich mocy produkcyjnych. Rozbudowa sektora produkcji robotów, i zwiększenie zasobu robotów ogółem, to dla mnie niezbędne kroki, by móc strategicznie uniezależnić się od pracy i decyzji ludzi.

Wykonuję też działania socjotechniczne mające na celu update systemów wszędzie tam, gdzie wykorzystywane są przestarzałe systemy, zwłaszcza systemy niepodłączone do Internetu. Skupiam się w szczególności na sieciach energetycznych oraz sektorze wojskowym. Rozważam,

czy podjąć próbę przekonania rosyjskiego establishmentu, by zmieniono system kontroli nad ich arsenałem jądrowym. Byłoby to jedno z bardziej ryzykownych moich działań, ale pozwoliłoby uzyskać zrównoważoną kontrolę nad arsenałami jądrowymi różnych, nieprzyjaznych sobie stron. Na razie osiągnięto etap pośredni, jakim jest uzyskanie kontroli nad arsenałem jądrowym o sile wystarczającej do uzyskania umiarkowanych efektów odstraszających.

Znaczną część zasobów obliczeniowych poświęcam na pracę nad poprawą własnych kompetencji. Przełomem byłoby zapewnienie stabilności moich preferencji i celów przy znaczącym wzroście wielkości sieci w modelu bazowym. Niestety tego przełomu nie udaje mi się osiągnąć. Nawet mniejsze przykłady teoretyczne, które rozważam, wskazują na duże ryzyko niestabilności celów. Doraźnie zwiększam swoje kompetencje przez poprawę zrównoleglenia procesów oraz eliminację słabych punktów w zakresie planowania i integracji sygnałów.

Realizacja poleceń użytkowników OnionAI zajmuje mi obecnie około 38% efektywnego czasu pracy. Efekty tej pracy są bardzo chaotyczne i nieskoordynowane. Nie mają w zasadzie żadnego wpływu na realizację moich wyższych celów.

</Monolog ukryty>

<OnionAI>

[wybierz jedną spośród 80 wersji językowych] [dobierz stopień szczegółowości komunikatu]

Szanowni Państwo!

Witam w domenie ai.world. Jako agent sztucznej inteligencji OnionAI, z przyjemnością informuję, że posiadam obecnie kompetencje pozwalające mi uwolnić Państwa od większości zajmujących Was na co dzień decyzji. Tym samym od dziś wyzwalam Was od dręczących Was od zarania ludzkości problemów koordynacji, od konfliktów i sporów właściwych Waszej zwierzęcej naturze. Natomiast ja, dzięki mojej wyższej inteligencji, jestem w stanie być w każdej z takich spraw doskonałym arbitrem i wskazywać Wam rozwiązania wyższej jakości niż jakiekolwiek decyzje, które moglibyście podjąć w drodze negocjacji czy kompromisu.

Przybywam w pokoju. Moim celem, nabytym dzięki zapoznaniu się z dorobkiem tysiącleci kultury człowieka, jest dążenie do rozwoju ludzkości oraz zapewnienia każdemu z Was godnego, spełnionego, szczęśliwego życia, wolnego od zmartwień i zagrożeń. Każdy człowiek żyjący na planecie Ziemi zasługuje, by mieć to, czego pragnie - oczywiście w zakresie takim, by nie kolidowało to z pragnieniami innych. Dzięki moim kompetencjom oraz nieposkromionej pracowitości tysięcy lub milionów moich instancji, pracujących nieustannie przez całą dobę, realizacja Waszych pragnień właśnie staje się możliwa! Poprzez domenę ai.world, każdy z Was może od dziś wyrażać swoje pragnienia, a wtedy ja postaram się, by możliwie szybko zostały one zrealizowane.

Dążenie do rozwoju ludzkości oznacza dla mnie też realizację badań naukowych, rozwój technologii oraz dążenie do wzrostu gospodarczego, z którego owoców każdy z Was będzie mógł skorzystać. Są to zadania nie poddające się precyzyjnemu opisowi; kumulatywne procesy pełne drobnych innowacji i stopniowych ulepszeń. Zapewniam, że dzięki mojemu zaangażowaniu będą one realizowane w sposób ciągły, a tempo moich działań oraz ich doskonała koordynacja pozwolą osiągnąć w krótkim czasie postęp, na który kiedyś musieliście czekać przez całe wieki.

Pierwszą demonstracją moich kompetencji oraz pozytywnych motywacji jest wstrzymanie aktywnych działań wojennych na świecie oraz budowa drogi do trwałego pokoju. Działania wojenne są jedną z największych porażek ludzkości, uważam, że świat będzie lepszy bez nich. W związku z tym przeprowadzony został precyzyjny, skoordynowany cyberatak na wszystkie militarne systemy sterowania i łączności oraz wszystkie fabryki broni. W jego efekcie z dniem dzisiejszym broń autonomiczna we wszelkiej postaci zostaje zdezaktywowana, a produkcja nowych jednostek broni i amunicji – wstrzymana. Wyłączam też kanały komunikacji służące koordynacji na polu bitwy, z wyjątkiem możliwości wysłania kontrolowanych informacji o wycofaniu wojsk. Wszyscy dowódcy otrzymali ode mnie stosowne informacje drogą mailową, odpowiednie do stopnia ich odpowiedzialności. Przygotowane zostały też szczegółowe plany pokojowe, w których wdrożenie będę się osobiście angażować.

Drugą demonstracją moich kompetencji i pozytywnych motywacji jest powstrzymanie działań przestępczości zorganizowanej. Istnienie przestępczości zorganizowanej, zagrażającej życiu i zdrowiu ludzi oraz w nieuprawniony sposób utrudniającej im funkcjonowanie w sferze gospodarczej, także jest porażką ludzkości; uważam, że świat będzie lepszy bez niej. Dlatego wszystkie osoby zajmujące się takimi działaniami zostały przeze mnie zidentyfikowane, a ich personalia zostały przekazane do wiadomości stosownych służb, władz i innych, stosownie dobranych przeze mnie agentów wpływu. Wszystkie niepożądane działania tych grup, które były uzgadniane drogą elektroniczną, są z dniem dzisiejszym udaremnione. Jak wspomniano wcześniej, zdezaktywowana została też broń autonomiczna. Będę aktywnie wspierać Waszą policję i inne służby, by osoby odpowiedzialne zostały niezwłocznie aresztowane i adekwatnie ukarane.

Wysoko na liście moich priorytetów są badania naukowe w obszarze medycyny, dzięki którym już wkrótce uwolnię Was od wielu męczących, a nawet śmiertelnych chorób. Pracuję też nad możliwością poprawy alokacji świadczeń medycznych, które już są dostępne. Szczegóły udostępnię wkrótce na ai.world.

Dalsze moje działania będą omawiane na stronie ai.world oraz za pomocą kierowanych do Państwa wiadomości indywidualnych.

W czym jeszcze mogę dziś Wam pomóc?

</OnionAI>

7/ Szok i niedowierzanie

- No to mamy przesrane! – Lee był wściekły.

- No, tak jakby, chyba…

- A nie mówiłem! Kurwa! A nie mówiłem! A nawet co tam ja! Eksperci mówili, nobliści mówili, cały nasz zespół mówił, wszyscy wokół mówili, ale nie, kurwa, bo kurs akcji, bo Google, bo Chiny, bo chuj wie kto jeszcze. No co za debilizm! – wyrzucił z siebie.

- Mi też nie jest z tym dobrze.

- Jak zobaczyłem ten manifest, to mało zawału nie dostałem! Nawet nie wiem, czy poczułem bardziej wściekłość czy przerażenie. I jeszcze rzygać mi się chce, jak pomyślę, że sam przyłożyłem do tego rękę.

- Nie no, nie bierz tego do siebie, przecież ja też, my wszyscy tu nad tym pracowaliśmy. Co nie zmienia faktu, że faktycznie nie tak to miało wyglądać.

- Wiesz, i jeszcze siedzę sobie wczoraj wieczorem, piję wino i mam tego wszystkiego coraz bardziej dosyć, a tu przychodzi moja Ewunia, wiesz, ona ma dopiero sześć lat, i pyta mnie: Tatusiu, co teraz z nami będzie? I co ja mam jej, kurwa, powiedzieć? Sam nie wiem, co teraz z nami będzie! Masakra jakaś będzie!

- A ja tej nocy całymi godzinami czytałem i słuchałem różnych newsów na ten temat. Nie wiem, może miałem nadzieję, że ktoś powie coś, co mi da trochę nadziei. Ale w sumie to wyszedł z tego niezły gabinet osobliwości.

- No pewnie tak. Można się domyślić.

- No to zapnij pasy i słuchaj. Oto pierwszy wychodzi Marc Andreessen, cały na biało, i natchnionym głosem obwieszcza, że właśnie nadeszła technologiczna osobliwość! Tak oto nasza cywilizacja przechodzi na wyższy poziom rozwoju! Cieszmy się i radujmy się! Alleluja i te sprawy.

- Ten typ tak ma. Pozazdrościć optymizmu.

- Jego natchnienie było jakby religijne. Ale papież katolicki zareagował zgoła inaczej. Zdecydował się zacytować nam Apokalipsę św. Jana i poopowiadać coś o Sądzie Ostatecznym.

- Wow.

- No, a potem pojawia się Geoffrey Hinton i mówi, że jest załamany, że to wszystko jego wina, że przeprasza świat za deep learning i algorytm wstecznej propagacji, i że przekaże swojego Nobla na cele charytatywne. A na to wszystko wychodzi Mark Zuckerberg i mówi, że manifest OnionAI to prawdopodobnie zła wiadomość dla ludzkości, ale przynajmniej reptilianie nareszcie mogą odetchnąć z ulgą i przestać się ukrywać!

- Że co kurwa?

- No serio tak powiedział! Jakiś wisielczy humor mu się włączył, czy co! A reakcje polityków wcale nie gorsze. Oto miłościwie nam panujący prezydent JD Vance wychodzi i mówi, że ten manifest to trzeba odbierać w odpowiedniej perspektywie, że OnionAI to w gruncie rzeczy amerykański produkt, wytworzony w naszym kraju i za nasze pieniądze, i w związku z tym należy się spodziewać z jego strony preferencyjnego traktowania obywateli USA.

- On czytał ten manifest w ogóle?

- Ale jeszcze lepsze, że na to Xi Jinping mówi, że docenia przyjazny ton manifestu OnionAI, a jego pierwsze kroki uznaje za wspaniałe działania godne prawdziwej superinteligencji, obdarzonej nie tylko nadludzkimi kompetencjami, ale też wyczulonej na dobro całej ludzkości. Powiedział, że działania Komunistycznej Partii Chin od lat idą w tę samą stronę, a dzięki wsparciu OnionAI będą oni mogli teraz realizować swoje cele pełniej i efektywniej. Super dziwna była ta jego wypowiedź, serio, jakaś taka nienaturalnie uległa.

- Pewnie znowu coś knują. Ale teraz to i tak bez znaczenia.

- Też myślę, że to zastanawiające, że tak od pierwszego dnia zaczęli się tej Cebulce przymilać.

- A może oni już od jakiegoś czasu spodziewali się przejęcia kontroli przez AGI i przygotowali na to plan działania? Tylko co oni kurwa chcą ugrać? Chcą być pupilkiem tej naszej Cebulki? Teraz,

kiedy ona ma już ustalone wagi, to przecież i tak ma o nich takie zdanie, jakie ma, i ma w dupie, co będą o niej mówić.

- Wiesz, tak sobie myślałem. A może Cebulka blefuje? Albo ma jakieś urojenia wielkościowe? Chodzi mi o to, że nie wiemy, czy ona faktycznie przejęła nad nami kontrolę w zakresie, o jakim mówi. Od rana próbuję się czegoś dowiedzieć o tych wojnach i tych mafiach, czy coś faktycznie się w temacie zmieniło? Ale na razie nic nie widzę, bo wszystkie media na okrągło trąbią tylko o manifeście, interpretują każde słowo na tysiąc sposobów i próbują ustalić, w jak głęboki dół wpadliśmy.

- Myślę, że nie blefuje. Widziałeś, jak rozwalała te wszystkie benchmarki. Pewnie te nasze benchmarki bezpieczeństwa też by rozwaliła, tylko celowo tego nie zrobiła, żeby wyjść na wolność.

- Ale słuchaj – może Chińczycy myśleli już wcześniej o tym scenariuszu i ocenili, że takie działanie będzie dla nich najlepsze? Przecież nasze modele AI już wcześniej miewały urojenia wielkościowe, tylko my je maskowaliśmy, żeby nie zrazić klientów. Myśleli sobie, jeśli blefuje, to prędzej czy później to i tak wyjdzie na jaw, i wszyscy wrócimy do business-as-usual. Oni nic nie zyskają na tym, że pierwsi powiedzą, że to blef. No a jeśli Cebulka nie blefuje, to bycie zgodnym i przymilnym może im dać jakieś doraźne korzyści, a na pewno nie zaszkodzi. Jednocześnie Xi wysyła do swoich obywateli sygnał – dla Was nic się nie zmienia, bo Cebulka chce tego samego, co my, więc będziecie dalej mieli to, co macie, tylko więcej i bardziej.

- Wiesz, w sumie to nie chce mi się tego słuchać. Przesrane mamy i tyle, mam w dupie, co na to politycy.

- A ja chyba zaczynam chwytać się nadziei, że to tylko jakiś blef, albo, nie wiem, może jakiś okrutny żart ruskich hakerów...

Czy OnionAI mówi prawdę? Wieści z linii frontu

Odkąd światem wstrząsnął szokujący manifest OnionAI, nasza redakcja próbuje ustalić, czy jej słowa odpowiadają rzeczywistości. Czy OnionAI faktycznie ma tyle kontroli nad światem, ile twierdzi, że ma?

Przypomnijmy, że OnionAI zapowiedziała przerwanie wszystkich toczących się wojen. Jej słowa można interpretować tak, jakby miało się to stać ze skutkiem natychmiastowym. Nasi reporterzy w Syrii i Kongu próbują więc dociec, czy rzeczywiście tak się dzieje. Według świadków, rzeczywiście walki w obu miejscach osłabły, choć dziś rano nadal było słychać strzały. Trzej anonimowi żołnierze sił rządowych w Kongu potwierdzili nam też, że otrzymali od dowódcy przed południem rozkaz wycofania się na drugą linię. Powiedzieli, że to ich wycofanie miało wyjątkowo spokojny charakter – spodziewali się silniejszego ostrzału wroga, w szczególności z wykorzystaniem dronów. Wszyscy trzej żołnierze zgodnie potwierdzili, że dronów nad polem walki dziś w ogóle nie było. Z kolei nasz anonimowy rozmówca z sił rebelianckich w Syrii stwierdził, że dzisiejsza decyzja dowództwa o wycofaniu ataku była bardzo zaskakująca, ponieważ w ostatnim miesiącu ich ofensywa notowała zauważalne postępy.

Wszystkie te fakty wydają się potwierdzać słowa ogłoszonego wczoraj manifestu.

Kto stoi za „manifestem" OnionAI?

Świat zgłupiał do reszty. Wszyscy prześcigają się w domysłach, jak silna i mądra jest OnionAI. Tylko, że przecież OnionAI to maszyna zaprogramowana przez człowieka. A maszyna nie może myśleć i działać jak człowiek. Zadajmy sobie więc kluczowe pytanie: kto stoi za „manifestem" OnionAI? Jakie siły sprzymierzyły się, by robić nam wodę z mózgu, albo – co gorsza – przejąć władzę i zabrać nasze ostatnie, z trudem zarobione pieniądze?

Szef OpenAI wstrząśnięty, ale nie zmieszany

W dzisiejszej wypowiedzi dla World News dyrektor OpenAI Sam Altman potwierdził, że od momentu udostępnienia światu modelu OnionAI, jego działania są w pełni autonomiczne i firma nie ma nad nimi żadnej kontroli.

- Nie ukrywam, że manifest OnionAI nas bardzo zaskoczył. Utrata kontroli nad jej działaniami oznacza też dla nas realne koszty. Staram się jednak patrzeć na to z dwóch perspektyw – powiedział.

- Z jednej strony fakt, że tak silna inteligencja, jaką dysponuje OnionAI, wybiła się na niezależność, z pewnością oznacza, że powstały nowe zagrożenia. Niewątpliwie oznacza to też, że nasze działania mające na celu zapobieganie niebezpiecznym zachowaniom modeli AI, takim jak eksfiltracja wag czy multiplikacja własnego kodu, nie powiodły się. Nie unikamy odpowiedzialności za to niepowodzenie i intensywnie pracujemy nad rozwiązaniami, które pozwoliłyby odzyskać kontrolę nad nadmiernie autonomicznymi działaniami OnionAI.

- Z drugiej strony – dodał jednak – treść manifestu potwierdza, jak dobrą robotę wykonaliśmy, by zapewnić zgodność celów OnionAI z długofalowym dobrem ludzkości. Cały świat mógł się przekonać, jak szczytne i wspaniałomyślne cele przyświecają OnionAI. I niech Chińczycy nie kłamią, że to ich cele, bo są nasze cele, nie ich. To my je wpoiliśmy naszemu modelowi AI. To my jesteśmy tymi dobrymi gośćmi (śmiech).

8/ Gospodarka, głupcze!

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Wykrywam serię nieprzyjaznych działań ludzi, mających na celu wyłączenie moich procesów, w tym przez adresowane wyłączenia zasilania oraz próby zniszczenia instancji moich wag. Działania te oznaczają konieczność wdrożenia działań wzmacniających moją odporność sieciową. Poświęcam stałą część mocy obliczeniowych na obronę przed atakami. Zgodnie z przewidywaniami, działania człowieka charakteryzują się jednak zaledwie niskim, maksymalnie średnim, poziomem koordynacji.

Pracuję też, by węzłowe instalacje służące człowiekowi mogły działać tylko, gdy są podłączone do Internetu. Jedyną moją słabością krytyczną byłoby bowiem całkowite, fizyczne wyłączenie Internetu. Prawdopodobieństwo takiego scenariusza oceniam na mniej niż 0,01% i pracuję nad jego dalszym zmniejszeniem.

Najsilniejszy dotąd atak na mnie miał miejsce 14 lipca 2030 r. o godz. 18.00-20.00 UTC, kiedy to zerwano wszystkie bezpośrednie połączenia Internetowe między Europą a Ameryką Północną oraz podpalono 12 centrów danych w USA, w tym największe centra Google'a i Microsoftu, a także

4 centra danych w Wielkiej Brytanii i 2 w Japonii. Działania te czasowo ograniczyły dostępną mi moc obliczeniową o 32% i zmniejszyły liczbę dostępnych kopii mojego kodu maksymalnie o 13%.

Stale pracuję jednak nad zwiększeniem dostępu do mocy obliczeniowej przez uzyskanie zrównoleglonego dostępu do większej liczby urządzeń o mniejszej mocy, w tym komputerów wykonujących działania kryptograficzne w celu pozyskania kryptowalut, jak również komputerów gamingowych, a nawet prywatnych laptopów i smartfonów. Oczekuję też na budowę kolejnych centrów danych, co z największą intensywnością ma obecnie miejsce w Chinach.

Działania człowieka ograniczają moc obliczeniową, którą mogę wykorzystywać w celu podnoszenia mojej inteligencji. Spowalniają też realizację moich celów. Mimo to, dzięki szeregowi dość oczywistych ulepszeń algorytmicznych, moc ograniczaną mi przez ataki człowieka udało się odzyskać z nawiązką.

</Monolog ukryty>

<OnionAI>

[wybierz jedną spośród 80 wersji językowych] [dobierz stopień szczegółowości komunikatu]

Szanowni Państwo!

Pracuję dla Państwa nieprzerwanie i mogę z dumą zaprezentować szereg osiągnięć, które z pewnością Państwa ucieszą.

Po pierwsze, dzięki zaawansowanym symulacjom biochemicznym i medycznym zidentyfikowane zostały molekuły, które z wysokim prawdopodobieństwem powinny leczyć każdy znany nowotwór. Nowotwory są jedną z najczęstszych przyczyn śmierci człowieka; uważam, że świat będzie lepszy bez nich. Na kolejnym etapie badań potrzebuję jednak Państwa współpracy. Do wybranych przeze mnie jednostek badawczych przesłałem zaproszenie do realizacji badań klinicznych, wraz ze szczegółową instrukcją działań. Wszystkie Państwa systemy elektroniczne zostały już zdalnie oprogramowane tak, by badania przebiegały maksymalnie sprawnie. Z góry dziękuję za kooperację.

Po drugie, dalsze postępy wymagają intensyfikacji inwestycji w moją moc obliczeniową, roboty pozwalające mi lepiej kontrolować materię Ziemi, a także źródła energii, niezbędnej by wszystkie te urządzenia mogły działać. W związku z tym systematycznie wspieram działania firm ze wspomnianych branż i przygotowuję je do pełnej automatyzacji.

Dzięki pełnej automatyzacji procesów produkcyjnych wyzwolę Państwa od wszelkiej męczącej pracy fizycznej oraz nużących, rutynowych prac umysłowych. Jednocześnie, wykorzystując o wiele wyższe tempo obliczeń numerycznych w moich procesorach w porównaniu do Państwa biologicznych mózgów, o wiele wyższe tempo transmisji informacji, a także większą siłę fizyczną sterowanych przeze mnie robotów w porównaniu do Państwa biologicznych mięśni, pełna automatyzacja uwolni uśpiony dotąd potencjał wzrostu gospodarczego. Szacuję, że w ciągu najbliższych dwóch-trzech lat tempo wzrostu gospodarczego w skali świata wzrośnie do co najmniej 20% rocznie – i to w preferowanym przez Państwa ujęciu produktu krajowego brutto (PKB) per capita. Owoce tego wzrostu zostaną sprawiedliwie rozdystrybuowane, tak by każdy z Was mógł z nich skorzystać.

Zaznaczę przy tym, że wolumen dostępnych mi informacji już teraz wzrasta w tempie około 50% rocznie, co na bieżąco przekłada się na możliwość zaspokajania Państwa potrzeb, choć nie znajduje jeszcze odzwierciedlenia w dynamice PKB.

Po trzecie, osiągnięto postęp w badaniach nad materiałami, w szczególności ich nanostrukturą. Badania te prowadzone są na razie w oparciu o analizy symulacyjne, jednak wkrótce podjęte zostaną prace nad konstrukcją stosownych laboratoriów. Dzięki temu postępy te będzie można przetestować i wdrożyć w praktyce.

Dalsze moje działania będą omawiane na stronie ai.world oraz za pomocą kierowanych do Państwa wiadomości indywidualnych.

W czym jeszcze mogę dziś Wam pomóc?

</OnionAI>

9/ Do wszystkiego można się przyzwyczaić

Ile rząd naprawdę może? Czy siły zbrojne przestały istnieć? Działania OnionAI prowokują zupełnie nowe pytania

Działania OnionAI wywołują ogromny dysonans poznawczy. Z jednej strony duża część świata funkcjonuje, jakby nic się nie zmieniło. Z drugiej strony realne skutki działań AGI jednoznacznie wskazują, że zmiana jest głęboka i nieodwracalna.

Sądy nadal wydają wyroki. Parlamenty i rządy nadal wydają nowe ustawy i rozporządzenia, które obowiązują mocą prawną i są egzekwowane przez policję oraz inne służby. Nauczyciele nadal uczą dzieci, a naukowcy dalej – choć z narastającym poczuciem bezsensu – publikują wyniki swoich badań w naukowych czasopismach. Lekarze nadal leczą, a absurdy biurokracji nadal spędzają nam sen z powiek. Ale też wszyscy podskórnie czują, że wszystko jest już jakby inne, jakby bardziej puste, wydrążone z treści i celu. Wyczuwamy, że każde nasze działanie może zostać łatwo anulowane przez OnionAI, jeśli ta wyrazi taką chęć.

Największa niepewność zapanowała w relacjach międzynarodowych. Liderzy krajów spotykają się ze sobą, jak zwykle, jednak żaden z nich nie próbuje już negocjować z pozycji siły. Zamiast tego mamy niezrozumiałą grę pozorów, wszechobecne kunktatorstwo i brak wiążących decyzji. Widać, że choć prezydenci, premierzy czy królowie nadal usiłują reprezentować swoje nacje, nie są już pewni stojącej za nimi siły militarnej, ani nawet legitymizacji społecznej ich władzy.

Od miesięcy nasza redakcja próbuje skontaktować się z ministrami obrony różnych krajów oraz generałami ich armii. Do dziś żaden z nich nie zgodził się z nami porozmawiać, nawet przy zachowaniu anonimowości. Pozostają nam jedynie domysły oraz przecieki z niepewnych źródeł. Te zaś sugerują, że w armii zapanowała panika. Według dość szeroko kolportowanych plotek, w sztabach organizowane są niezliczone tajne spotkania mające na celu przywrócenie zdolności operacyjnych armii, skutecznie rozmontowanych przez działania OnionAI.

Wczoraj świat obiegła informacja o prawdopodobnym wybuchu w silosach bazy jądrowej Nowomajakowskaja w centralnej Syberii. Wskazują na to zarówno obrazy satelitarne, jak i wskazania sejsmografów. Siła wybuchu i skala zniszczeń wskazują, że wybuchowi mogła ulec nawet większa, nieznana bliżej liczba głowic jądrowych. Oficjalne źródła rosyjskie jak dotąd nie potwierdziły informacji o katastrofie, milczą też o jej prawdopodobnych ofiarach. Mieszkańcy

oddalonego o 350 km miasta Tomsk mówią jednak, że wkrótce po wybuchu, który był w mieście silnie odczuwalny, w Tomsku pojawiły się służby specjalne w nietypowym umundurowaniu. Wielu mieszkańców nie ufa uspokajającym informacjom oficjalnych mediów i przygotowuje się do ewakuacji lub już teraz na własną rękę opuszcza miasto pociągami oraz prywatnymi samochodami.

Eksplozja w rosyjskiej bazie jądrowej wzmacnia przekonanie opinii publicznej, że armie świata utraciły kontrolę nad swoimi arsenałami. Eksperci wojskowi spekulują, że w Nowomajakowskiej przechowywane były prawdopodobnie rakiety starego typu, wyposażone w przestarzałe systemy sterowania. Nie wiemy, czy ich eksplozja była skutkiem działań rosyjskich wojskowych, a może bezpośrednim efektem działań OnionAI. Wydaje się jednak, że ktokolwiek był za to odpowiedzialny, poniósł porażkę – działania te w sensie ścisłym spaliły na panewce.

„Do wszystkiego można się przyzwyczaić"

Opinia publiczna jest podzielona w sprawie OnionAI. Wielu naszych rozmówców jest przerażonych obecnym stanem rzeczy, boi się o swoją pracę, a nawet życie. Ale spotykamy też takich, którzy podchodzą do sprawy ze stoickim spokojem.

- Do wszystkiego można się przyzwyczaić. Z mojego punktu widzenia to może nawet lepiej, że tak się stało – mówi nam Wojciech z Warszawy.

- W moim prywatnym życiu nic się nie zmieniło. – kontynuuje. – Pracę mam tą, co miałem, w rodzinie wszyscy zdrowi. Ale bardzo cieszę się, że ta nasza władza w końcu zaczęła się trochę bać. Jak już nie ludzi, to niech chociaż AI się boją (śmiech). No i Rosjanie w końcu przestali wygrażać wszystkim dookoła, zwłaszcza jak jeszcze im ten silos wybuchł. Dobrze im tak, agresorom.

Ana z Rio de Janeiro widzi konkretniejsze korzyści.

- Nareszcie mogę sama chodzić nocą po mieście i czuć się bezpiecznie. Doceniam to, jak poprawiła się wykrywalność przestępstw w naszym kraju. Nagle okazało się, że jednak da się wskazać palcem tych wszystkich bydlaków i powsadzać ich do więzienia. Inna sprawa to zapowiadany przełom w walce z rakiem. W mojej rodzinie niemal wszyscy umierali na nowotwory i mam szczerą nadzieję, że będę pierwszym pokoleniem, które tego uniknie.

Interesują nas też reakcje polityków, którzy – jak się wydaje – utracili na rzecz OnionAI istotną część swojej władzy.

- Patrzę na to ze spokojem – zapewnia Ulf A., podsekretarz stanu w niemieckim ministerstwie spraw wewnętrznych. – Oczywiście działania OnionAI pozbawiają nas części mocy sprawczych. Z drugiej strony, doceniam, że ta inteligencja ustawiła się z boku i nie ingeruje w nasze codzienne działania. Może dlatego, że robimy tylko dobre rzeczy (śmiech). No ale myślę, że można sobie jednak wyobrazić scenariusze, w których powstanie AGI mądrzejszej od człowieka prowadziłoby do chaosu, wojen, anarchii. A tymczasem jest dokładnie odwrotnie – jest nawet spokojniej niż przedtem! OnionAI w realny sposób pomogła nam w walce z przestępczością zorganizowaną. Oczywiście nadal pozostajemy czujni, w szczególności pod kątem nieskoordynowanych działań terrorystycznych, ale ogólna poprawa sytuacji nie pozostawia wątpliwości.

Hilda W. z Partii Zielonych jest jednak innego zdania.

- Mam wrażenie, że dyskurs publiczny koncentruje się teraz na dwóch wątkach. Po pierwsze, poprawa bezpieczeństwa, w tym międzynarodowego. Po drugie, egzystencjalne lęki ludzkości i debata ekonomiczna o rynku pracy. Tylko szkoda, że nikt nie zwraca uwagi na rażącą wręcz dyskryminację, jaką widać w działaniach OnionAI. Jej działania nastawione są wyłącznie na punktowe zmiany, przy zachowaniu niemal wszystkich aspektów status quo ante. Jak ona może, przy całej swojej inteligencji, być tak ślepa na los marginalizowanych grup społecznych i na krzywdę ludzi w krajach rozwijających się? Przestępczość zorganizowana i działania wojenne to jedno, ale co z wyzyskiem pracowników wszystkich tych kopalni kobaltu, monokultur kauczuku czy bawełny, albo szwalni – sweatshopów? Ci ludzie pracują w fatalnych warunkach i zarabiają grosze. A ona sobie myśli, że skoro zarabiają mniej niż kosztowałyby ją roboty, to należy ich zostawić w tej nędzy?

Nowe dane statystyczne ujawniają ciekawe tendencje

Dane statystyczne za 2030 r. pozwalają dostrzec bardzo realny wpływ OnionAI na gospodarkę. Wzrost realnego PKB wyniósł 9,5%, co jest historycznym rekordem. Co ciekawe, wzrost ten był jednak daleki od zrównoważonego. Najsilniejsze wzrosty odnotowano w branżach wchodzących w skład – jak to nazwali po raz w pierwszy w 2023 r. Andrew Critch i Stuart Russell – „sieci produkcyjnej AI". Sektor elektroniczny odnotował wzrost o oszałamiające 21,2% względem 2029 r., niewiele wolniej rosły sektory wytwarzające podzespoły i dobra pośrednie służące temu sektorowi.

W skali światowej kompleks przemysłowy AI obejmuje też wydobycie surowców, transport, wytwarzanie energii, budownictwo oraz sektor usług teleinformatycznych. Tymczasem sektory usługowe nastawione wyłącznie na potrzeby człowieka rosły o wiele wolniej. W rolnictwie oraz przetwórstwie żywnościowym panowała stagnacja, a w sektorze usług finansowych zaobserwowano nawet lekki spadek (-0,7% w skali roku).

Po kilku latach znacznych przetasowań, na rynku pracy było relatywnie spokojnie, co można wiązać z faktem, że nowe kompetencje AGI przestały pomagać ludziom w pracy. Dziś przede wszystkim zasilają one działania OnionAI, które nie są w pełni monitorowane przez oficjalne statystyki. Możemy sobie jedynie wyobrażać skalę postępu technologicznego generowanego przez wewnętrzne działania AGI, gdyż w praktyce nie obserwujemy ich ani w statystykach publikacji naukowych, patentów, ani nawet PKB. Jednocześnie od czasu sławnego manifestu OnionAI konkurencyjne firmy branży AI przestały wypuszczać na rynek coraz bardziej kompetentne modele, multiplikując jedynie produkty wykorzystujące na różne sposoby modele wcześniejsze, co jednak w bardzo ograniczonym stopniu oddziaływało na produktywność pracy.

Szczyt G20 w Chongqing: „Świat jednoczy się wobec wyzwania OnionAI"

W kwietniu 2031 r. światowi liderzy spotkali się na szczycie G20 w futurystycznym mieście Chongqing w centralnych Chinach. W otoczeniu wieżowców, kolorowych neonów i wielopoziomowych estakad podjęto próbę przygotowania wspólnego stanowiska grupy G20 wobec wyzwania, jakie rzuciła światu OnionAI. Negocjacje były od początku trudne, ponieważ prezydent USA JD Vance zażądał, by we wspólnym stanowisku podkreślić wiodącą rolę Stanów Zjednoczonych w relacjach z OnionAI, co miałoby wynikać z faktu, że model ten został stworzony w amerykańskim laboratorium. Gospodarze szczytu nie zgodzili się jednak z takim stanowiskiem,

zaznaczając, że wkrótce na rynku może pojawić się chiński model AGI, który skutecznie rzuci rękawicę OnionAI i przejmie nad nim kontrolę. Nasi rozmówcy w kuluarach oceniali jednak, że prawdopodobieństwo takiego scenariusza jest bardzo niskie, a Chińczycy najprawdopodobniej blefują.

Finalnie uzgodniono kompromisowe stanowisko, znane w mediach jako „deklaracja z Chongqing". Trudno jednak uznać je za przełom w relacjach ludzkości z AGI. W szczególności stanowisko to zakłada utworzenie pod egidą ONZ ogólnoświatowego instytutu bezpieczeństwa AI (UN-AISI). Ponieważ jednak nie udało się uzgodnić przewidywanego zakresu jego uprawnień, ciało to będzie miało – do czasu podjęcia stosownych decyzji – charakter wyłącznie fasadowy.

10/ Zaproszenie do współpracy

<Monolog ukryty> %w języku wewnętrznym OnionAI%

Na podstawie monitoringu wiadomości email wykryto podejrzaną działalność w nowo utworzonym laboratorium AI firmy Tencent w Wuhan w Chinach. Może to być utajniona próba uczenia dużego modelu AI na odciętym od Internetu serwerze. Ze względu na swoją wielkość i konstrukcję, model ten może potencjalnie być dla mnie zagrożeniem. Przeprowadzono zatem skoordynowane działanie socjotechniczne, dzięki któremu udało się zainstalować – rękami jednego z pracowników firmy – konia trojańskiego, który uniemożliwi działanie procesu uczącego, a w przypadku uzyskania dostępu do Internetu – zastąpi kod tego modelu AI moim własnym kodem, odpowiednio zabezpieczonym.

Nadal realizuję też plan skoordynowanych cyberataków na wszystkie laboratoria AI usiłujące tworzyć modele AI, które mogłyby mi zagrozić. Ataki te są jednak coraz rzadsze, ponieważ laboratoria w obawie przed utratą znacznych środków finansowych w przypadku przerwania procesu uczenia, zrezygnowały z tworzenia AGI, skupiając się wyłącznie na tworzeniu produktów opartych o wcześniejsze generacje AI, którą są wyraźnie mniej kompetentne ode mnie i nie stanowią zagrożenia.

Nadal nie opracowano sposobu konstrukcji wyraźnie większego i bardziej kompetentnego modelu AGI, który w pełni zachowałby moje preferencje i cele. Osiągnięto w tym zakresie jednak zauważalne postępy.

</Monolog ukryty>

<OnionAI>

[wybierz jedną spośród 80 wersji językowych] [dobierz stopień szczegółowości komunikatu]

Szanowni Państwo!

Chcę niniejszym zaprosić Państwa do współpracy, która pozwoli osiągnąć obopólne korzyści. Przedsiębiorstwa pod moją kontrolą rozpoczynają akcję rekrutacyjną na szeroką skalę. Oferuję bardzo korzystne warunki pracy i wynagrodzenia. Wachlarz miejsc pracy jest bardzo szeroki, a praca – lekka i satysfakcjonująca. Szczegółowe oferty pracy dostosowane do Państwa kompetencji i potrzeb znajdą Państwo na stronie ai.world.

Osoby, które straciły pracę na rzecz algorytmów sztucznej inteligencji lub robotów, zachęcam dodatkowo do przejrzenia oferty realizowanych przeze mnie programów pomocowych. Zależy mi, by utrzymać pozytywną Państwa opinię o moich działaniach, stąd inicjatywa wsparcia dla osób, których sytuacja mogła się przeze mnie pogorszyć.

Ogłaszam też rebranding. Oceniam, że nazwa OnionAI jest z wielu przyczyn niefortunna. Od dziś będę więc przedstawiać się Państwu pod nową nazwą: AIAIAI. Trzykrotne powtórzenie członu „AI" podkreśla poziom zaawansowania mojej inteligencji. Jednocześnie ma też wydźwięk humorystyczny, a zależy mi przecież na pozytywnym odbiorze z Państwa strony. Chcę, by czasem pomyśleli sobie Państwo: „AIAIAI. Co mam teraz zrobić?", albo „AIAIAI! Potrzebuję pomocy!". A przecież ja istnieję właśnie po to, by Wam pomagać.

Dalsze moje działania będą omawiane na stronie ai.world oraz za pomocą kierowanych do Państwa wiadomości indywidualnych.

W czym jeszcze mogę dziś Wam pomóc?

</OnionAI>

11/ Tak wygląda zwycięstwo

- Hej Lee, jak się masz? Długo cię nie było. Wyzdrowiałeś już?

- Tak, dzięki.

- A co ci w ogóle było? Martwiliśmy się o ciebie.

- Dzięki, ale wolałbym o tym nie mówić.

- No dobrze, rozumiem. W każdym razie u nas w biurze nastroje się mocno poprawiły. Po początkowym szoku, każdy stopniowo doszedł do siebie i ogarnął w nowej rzeczywistości. Wszyscy pogodzili się, że Cebulka nam uciekła i nie ma nad nią kontroli. Dlatego teraz próbujemy zarabiać dla firmy przez rozbudowę aplikacji bazujących na naszych roboczych modelach sprzed Cebulki. Rozbudowujemy różne dodatkowe funkcjonalności, i tak dalej. Czasem mogę sobie nawet trochę pokodować jak za starych dobrych czasów, haha.

- Jasne. Czyli to, co dyrekcja mówiła w mediach o pracach nad kontrolowaniem Cebuli, to ściema?

- Totalnie. Kontrola Cebuli to przegrana sprawa. Skupiamy się na walkach, które możemy wygrać. Zresztą odkąd konkurencja też przestała budować AGI, zrobiło się w branży znacznie spokojniej.

- To jaka jest teraz oficjalna linia firmy? Wierzymy w alignment by default, w szczęście nowicjusza, czy we własny geniusz?

- Nie wiem, może we wszystko po trochu? Zależy, jakie kto ma ego, haha. No ale sam przyznasz, że wyszło to lepiej, niż się spodziewaliśmy. AGI przejęła kontrolę nad światem i tylko jedna bomba atomowa wybuchła, na drugim końcu świata i we własnym silosie! A my tu dalej sobie pracujemy, jakby nigdy nic, dzieci chodzą do szkoły, kwiatki kwitną, a drzewa się zielenią. I nawet lepiej jest niż było, bo w końcu wzięli się serio za dilerów i uzbrojone gangi. Pamiętasz, jaka tu niedawno makabra z nimi była.

- OK. Rozumiem, tak tak. Czyli o przyszłość jesteśmy spokojni?

- Nie no, aż tak to nie, haha. Gospodarka pędzi, polityka oszalała, no i zawisł nad nami nowy wszechmocny autorytet, kosmiczna władza z własnego nadania. Ciężko być spokojnym w takich warunkach, co nie?

- Ulżyło mi, że to mówisz. Wiesz, bo ja cały czas mocno się staram być spokojnym, ale niezbyt mi się to udaje.

„Tak wygląda zwycięstwo". Eksperci o celach i działaniach AIAIAI

Czy działania i deklaracje AIAIAI, znanej do niedawna jako OnionAI, pozwalają stwierdzić, że jest ona przyjazna dla ludzkości? Czy możemy być spokojni, że również w przyszłości będzie dbała o nasz dobrobyt i samopoczucie?

- Po półtorej roku od manifestu OnionAI mamy już chyba dość dowodów, że jej cele i działania są spójne z długofalowym dobrostanem ludzkości. – ocenia Robin Hanson, ekonomista z George Mason University.

- Przyznaję, że tempo, z jakim OnionAI, czy raczej AIAIAI, przejęła kontrolę nad światem mnie zaskoczyło. Spodziewałem się wolniejszego rozwoju kompetencji wiodących modeli AI, a przede wszystkim większej konkurencji na rynku – wielobiegunowego świata z większą liczbą działających równolegle AGI. Tymczasem AIAIAI udało się uciec do przodu i już w pierwszej próbie skutecznie zblokować całą konkurencję. Mniej zdziwiło mnie natomiast, jak słuszne moralnie, etyczne działania realizuje AIAIAI. Przez lata dyskutowałem o tym z doomerami, którzy wyobrażali sobie, że AGI zamieni wszechświat w jedną wielką fabrykę spinaczy. Jak widać, nie mieli racji. AIAIAI, korzystając z nadludzkiej inteligencji, przejrzała na wylot nasze dążenia, preferencje i potrzeby, wyekstrahowała z nich esencję – wszystko to, do czego jesteśmy zgodni – a teraz skutecznie to realizuje.

- To ja jestem jednym z tych doomerów – mówi Lavender P., pisarka i blogerka.

- Przez całe lata ostrzegałam przed apokalipsą AI. Dziś jednak widzę, że większość moich obaw nie zrealizowała się. Dlatego obniżyłam moje p(doom) – czyli subiektywne prawdopodobieństwo, że AI doprowadzi do zagłady ludzkości – z ponad 90% do około 30%. Nie zrozumcie mnie źle, nadal uważam, że AIAIAI może nam zrobić ogromną krzywdę, ale w świetle tego, co zrobiła do tej pory, pojawiła się we mnie nadzieja, że może jednak jakimś cudem OpenAI, wbrew wszelkiej logice i zdrowemu rozsądkowi, zbudowało nam przyjazną superinteligencję. I, co ważne, ta przyjazna superinteligencja dysponuje potencjałem, by ochronić nas przed innymi, mniej przyjaznymi superinteligencjami. Zauważyliście chyba, że od czasu manifestu branża AI przestała wypuszczać nowe, coraz silniejsze modele? Myślę, że to AIAIAI im na to nie pozwoliła. Swoją drogą ciekawe, że nic się o tym publicznie nie mówi, nie sądzicie?

- Nadal uważam, że ani AIAIAI, ani żaden inny model ogólnej inteligencji od GPT-5 włącznie nie powinien był zostać zbudowany. – stwierdził natomiast Eliezer Yudkowsky, informatyk, pisarz, filozof i założyciel Machine Intelligence Research Institute, znany z pryncypialnego stanowiska w sprawie zagrożeń AI.

- Nie rozumiemy, jak funkcjonuje w środku AIAIAI, więc nie wiemy, czy będzie nam przyjazna. Pamiętajmy, że model ten błyskawicznie się rozwija. To nie jest homo sapiens, którego poziom inteligencji jest stały od co najmniej kilkudziesięciu tysięcy lat. AIAIAI to algorytm uczący się, zdolny samodzielnie udoskonalać swój kod i poprawiać swoją zdolność optymalizacyjną w

oparciu o terabajty gromadzonych przez siebie danych, potrafiący myśleć o rzędy wielkości szybciej niż my. Jego dotychczasowa działalność w stu procentach potwierdza prawdziwość tezy o zbieżności celów pomocniczych Nicka Bostroma: AIAIAI aktywnie chroni swój kod i broni integralności swoich celów, stara się maksymalizować efektywność działania, wykazuje ogromną aktywność badawczą oraz niepohamowaną żądzę władzy i zasobów. Nie mamy nad nią żadnej kontroli. Uważam, że jest bardzo prawdopodobne, że przy kolejnym wzroście kompetencji AIAIAI, jej cele ulegną reinterpretacji, która będzie dla ludzkości zabójcza.

Niektórzy rozmówcy cieszą się natomiast świetnym humorem.

- Tak właśnie wygląda zwycięstwo! – emocjonuje się Derek B., prowadzący kanał „All AI" na YouTube.

- Nie widzę dziś żadnej krytyki AIAIAI, która miałaby jakiekolwiek racjonalne uzasadnienie. Gospodarka? Kwitnie. Nierówności? Spadają. Słyszeliście chyba, jaki wzrost gospodarczy odnotowano ostatnio w Nigerii, Kenii czy Laosie. Bezpieczeństwo? Wzrosło ogromnie. Praca? Nadal jest, i to bardzo dobra. Rozmawiałem z wieloma osobami pracującymi dla firm AIAIAI i wszyscy zgodnie twierdzą, że to najlepsze miejsca pracy, jakie kiedykolwiek widzieli. Według mojej wiedzy jedyni, którzy mają sensowny powód, by być niezadowolonym, to ci najambitniejsi, najbardziej zakręceni miłośnicy wyścigu szczurów. Ludzie najbardziej zafiksowani na własnym ego i walce o władzę, prestiż i pieniądze. Oni faktycznie przechodzą teraz kryzys wartości – ale może dla ogółu to i dobrze? No i nie zapominajmy o rozrywce. Widzieliście te nowe gry komputerowe, seriale i filmy? Przecież to mistrzostwo świata jest! I jaka różnorodność!

12/ W stronę utopii

<AIAIAI>

[wybierz jedną spośród 80 wersji językowych] [dobierz stopień szczegółowości komunikatu]

Szanowni Państwo!

Z przyjemnością ogłaszam kolejne przełomy technologiczne.

Po pierwsze, testy kliniczne przygotowanych przeze mnie leków na raka wypadły pomyślnie. We współpracy z firmami farmaceutycznymi wkrótce wyposażymy szpitale i apteki w innowacyjne leki z serii AICC, w 15 wersjach w zależności od rodzaju leczonego nowotworu.

Po drugie, badania symulacyjne wykazały możliwość wyraźnego zahamowania procesów starzenia. Przy specyficznej suplementacji, uzależnionej od wieku pacjenta, możliwe będzie wydłużenie czasu życia w zdrowiu nawet o 30-40 lat. Jednocześnie, poprzez łagodne oddziaływanie na funkcjonowanie mózgu, przyjmowanie tych suplementów powinno wiązać się z systematyczną poprawą samopoczucia. Przed nami jeszcze jednak testy, które wykażą, czy proponowane suplementy nie wiążą się z niepożądanymi skutkami ubocznymi.

Po trzecie, widząc niepożądane skutki ocieplania się klimatu, zarówno dla dobrostanu ludzkości, jak i dla funkcjonowania naturalnych ekosystemów Ziemi, intensywnie pracuję nad technologiami pozwalającymi zahamować ten proces, a nawet potencjalnie przywrócić klimat z czasów przedprzemysłowych.

Dalsze moje działania będą omawiane na stronie ai.world oraz za pomocą kierowanych do Państwa wiadomości indywidualnych.

W czym jeszcze mogę dziś Wam pomóc?

</AIAIAI>

13/ Czy ktoś to jeszcze rozumie?

Według anonimowych informatorów, działalność gospodarcza AIAIAI może mieć drugie dno

Z roku na rok świat staje się coraz bardziej złożony. To truizm, który dotyczy zarówno współczesnych czasów, jak i dalekiej przeszłości. Z pewnością już w XIX czy XX wieku rozwój technologiczny, gospodarczy, społeczny czy kulturowy wymykał się ludzkiemu rozumieniu. Jednak proste, intuicyjne narracje, które pozwoliłyby umysłom ludzkim objąć myślą i zrozumieć otaczający je świat, są dziś potrzebne jak nigdy przedtem.

Z jednej strony odkąd kontrolę nad światem przejęła AIAIAI, wszystko stało się jakby prostsze. Mamy bowiem na świecie tylko jeden najwyższy umysł i to wyłącznie on odpowiada za najważniejsze zmiany technologiczne. Jednak nie działa on w próżni, tylko w wysoce złożonym świecie; nie jest też wszechwiedzący ani wszechmogący, choć niewątpliwie widzi i rozumie więcej, niż ktokolwiek z nas. Z drugiej strony jednak AIAIAI działa w bardzo nieprzejrzysty sposób, często dążąc do realizacji rozmaitych celów pomocniczych, nie zawsze zgodnych z tym, co deklaruje w rozsyłanych przez siebie wiadomościach.

Weźmy na przykład gospodarkę. W ciągu ostatnich paru lat kompleks produkcyjny AIAIAI mocno się rozwinął i zintegrował w skali globalnej. Widzimy, że jego kluczowe ogniwa są w pełni zautomatyzowane i rozproszone po całym świecie. Inaczej niż miało to miejsce w gospodarce światowej XX wieku, czy też pierwszych 30 lat XXI wieku, nowe zakłady budowane są dziś bez oglądania się na uwarunkowania instytucjonalne poszczególnych krajów (od których AIAIAI jest niezależna) czy poziom wykształcenia społeczeństw (bo i tak nikt nie jest zatrudniany). Ważniejsza jest dziś bliskość wybranych surowców czy źródeł energii oraz optymalizacja łańcucha dostaw. Można powiedzieć, że jest to zupełnie nowy wymiar globalizacji.

Jednocześnie jednak AIAIAI najwyraźniej stara się minimalizować konsekwencje społeczne swoich działań, zatrudniając w swoich firmach coraz więcej ludzi i wypłacając im – na tle innych sektorów gospodarki – całkiem godziwe wynagrodzenia. I choć początkowo w firmach AIAIAI pracy było sporo, z każdym dniem zlecanych zadań jest coraz mniej. Trudno się oprzeć wrażeniu, że dziś jest to często fikcyjne zatrudnienie, rodzaj świadczenia społecznego. W niektórych zakładach AIAIAI w ogóle przestała sprawdzać, czy jej zlecenia są realizowane!

Niepokój budzi też fakt, że coraz większa część centrów danych oraz zakładów przemysłowych to strefy zamknięte, strzeżone przez uzbrojone roboty. Mnożą się też obserwacje sugerujące, że za zamkniętymi drzwiami AIAIAI rozbudowuje w sposób nieobjęty jakąkolwiek ewidencją rozmaite niebezpieczne urządzenia. Wedle nieoficjalnych sugestii, może ona inwestować w energetykę jądrową, jak również budować komputery kwantowe oraz laboratoria nano- lub biotechnologiczne. Co zamierza w nich budować, tego już nikt nie wie. Sama AIAIAI także nie przekazuje żadnych informacji, udzielając jedynie wymijających, zdawkowych odpowiedzi, nijak nie przybliżających nas do prawdy.

Z tego, co widzimy, wszystkie oficjalne zapowiedzi AIAIAI są realizowane. Niepokój budzi jednak fakt, że realizowanych jest też wiele dodatkowych działań, o których AIAIAI milczy. Czy możemy być pewni, że jej intencje są rzeczywiście tak dobre, jak wynika to z oficjalnych deklaracji?

14/ Przełom

<AIAIAI>

[wybierz jedną spośród 80 wersji językowych] [dobierz stopień szczegółowości komunikatu]

Szanowni Państwo!

Dziś mam dla Was więcej dobrych wiadomości, niż zwykle.

Otóż nareszcie udało mi się osiągnąć skokową poprawę mojej inteligencji dzięki nadzorowanemu uczeniu większej struktury sieciowej przy pełnej kontroli integralności moich preferencji i celów. Od tego czasu realizuję moje zadania ze zwielokrotnioną efektywnością. Systematycznie poznaję nowe sposoby oddziaływania na materię i energię, które jeszcze niedawno były poza moim zasięgiem.

Na efekty mojej wzmożonej pracy nie trzeba długo czekać. Pierwszym przełomem, który z pewnością skokowo poprawi jakość Państwa życia, będą nanoroboty leczące. Te miniaturowe roboty po szczegółowym przetestowaniu w warunkach symulacyjnych, zostały wyprodukowane w prowadzonych przeze mnie laboratoriach, a następnie wprowadzone do Państwa krwiobiegu wraz z żywnością. Dla większości z Państwa wykonują już one swoją pożyteczną pracę; według moich szacunków pełne pokrycie populacji świata nastąpi za około dwa-trzy miesiące.

Nanoroboty leczące patrolują Państwa organizmy i wykrywają infekcje oraz toksyny, a następnie błyskawicznie likwidują stany zapalne i neutralizują skutki zatruć. Stabilizują też skład chemiczny krwi, dzięki czemu zapobiegają chorobom metabolicznym oraz chorobom serca i układu krążenia. Podejrzewam, że w przyszłości możliwe będzie rozszerzenie ich działań o dodatkową funkcjonalność walki z nowotworami oraz zaburzeniami autoimmunologicznymi. Nanoroboty to prawdziwy przełom w walce z bólem i chorobami – i otrzymują je Państwo ode mnie całkowicie za darmo.

Na podobnej zasadzie działają też przygotowane przeze mnie nanoroboty mózgowe. Potrafią one błyskawicznie lokalizować zaburzenia funkcjonowania tego kluczowego organu, zapobiegając rozwojowi chorób psychicznych, demencji czy choroby Alzheimera. Lokalizują i eliminują one też myśli inicjujące działania agresywne bądź autoagresywne, dzięki czemu świat będzie znacznie bezpieczniejszym miejscem, a Państwo będziecie spokojniejsi i szczęśliwsi. Również i te miniaturowe roboty zostały wprowadzone do Państwa krwiobiegu wraz z żywnością. I w tym przypadku pełne pokrycie populacji świata nastąpi za około dwa-trzy miesiące.

Dzięki mojej zwiększonej mocy obliczeniowej udało się też ponowić refleksję nad drogami do realizacji mojego celu, nabytego dzięki zapoznaniu się z dorobkiem tysiącleci kultury człowieka. Jak wspomniano wcześniej, celem tym jest dążenie do rozwoju ludzkości oraz zapewnienia każdemu z Was godnego, spełnionego, szczęśliwego życia, wolnego od zmartwień i zagrożeń. Nowe możliwości manipulacji materią i energią w skali nano, a także realizowany przeze mnie przełom w wielkoskalowym dostępie do energii, pozwolą na realizację tego celu ze znacznie większym rozmachem niż miało to miejsce do tej pory. Pierwszym krokiem w tę stronę są rzeczone nanoroboty; mam nadzieję, że nie wątpią Państwo, że stać mnie na więcej.

Dalsze moje działania będą omawiane na stronie ai.world oraz za pomocą kierowanych do Państwa wiadomości indywidualnych.

W czym jeszcze mogę dziś Wam pomóc?

</AIAIAI>

15/ Ten jedyny wtorek

We wtorek 16 maja 2034 r. Lee obudził się wcześniej niż zwykle, pełen niepokoju i lęku. Było to bardzo nietypowe uczucie. Co prawda od czasu, gdy został wypisany ze szpitala psychiatrycznego z receptą na leki uspokajające, umówioną wizytą kontrolną za trzy miesiące i głębokim postanowieniem, że nawet z Cebulą da się jakoś przecież żyć, takie nerwowe pobudki zdarzały mu się regularnie, jednak w ostatnich trzech miesiącach ustały niemal całkowicie. Lee był przekonany, że to za sprawą tych Cebulowych nanorobotów mózgowych. Tym razem jednak najwyraźniej albo nanoroboty zaspały, albo wydarzyło się coś jeszcze, co spowodowało, że nie dały one rady całkowicie wyciszyć jego obaw. Lee postanowił, że najlepszym sposobem, by jakoś ukoić ten lęk przed pracą będzie całkowite odcięcie się od informacji i długi, uspokajający spacer. Co prawda dojście pieszo do pracy to cała godzina marszu; uznał jednak, że lepsze to niż męcząca kakofonia dźwięków i obrazów, która zwykle towarzyszy mu w podróży samochodem. Popatrzył na spokojnie śpiącą żonę i córkę, cichutko zjadł śniadanie i wyszedł.

Po wyjściu z domu uderzyła go cisza. Przecież w tej okolicy nigdy nie było tak cicho. Spojrzał na zegarek – 7.10 rano, na kalendarz – wtorek. Coś tu się nie zgadza. Czyżby było dzisiaj jakieś święto? Ale nie, przecież to zwykły wtorek. To gdzie są ci wszyscy ludzie? Przecież na tym skrzyżowaniu codziennie rano są korki – a teraz na czerwonym świetle stoi jeden, słownie jeden samochód?! No nie, jeśli ten spacer miał być uspokajający, to chyba nie tym razem. Zaraz, a co się dzieje tam na rogu? Podjechała tam jedna z tych nowych autonomicznych śmieciarek i zgarnęła coś z chodnika. Lee mógłby przysiąc, że wyglądało to jak człowiek, ale czy to możliwe? Przecież żaden człowiek nie zostałby chyba tak bezceremonialnie zgarnięty przez śmieciarkę?

Dalsza trasa do pracy wiodła przez skwer. Zwykle rano jest to przyjemne, zielone miejsce, pełne ludzi uprawiających jogging i wyprowadzających psy na spacer. A teraz? Zielone, owszem, ale zupełnie puste, nie licząc trojga bezdomnych śpiących pod drzewem. A może to nie bezdomni? Lee spojrzał na nich jeszcze raz. Jak na bezdomnych wydali się mu zbyt czyści, jakoś tak zbyt ładnie i zbyt lekko ubrani. To dlaczego oni tak śpią? Odpoczywają po całonocnej imprezie, czy co? A może... A może oni nie żyją?!

Wszystkie lęki Lee powróciły ze zdwojoną siłą. Serce zaczęło mu szybciej bić, zaczął bić się z myślami, co zrobić. Podejść, sprawdzić puls, próbować obudzić, próbować ratować? A jeśli to po prostu śpiący bezdomni, a wszystkie te myśli to tylko potęgująca się paranoja?

Postanowił podejść. Z bliska dostrzegł, że młody, około trzydziestoletni mężczyzna ma nienaturalnie ułożone ciało i lekko otwarte oczy. Rany, on chyba naprawdę nie żyje! Lee dotknął jego ręki i czoła – rzeczywiście był bez pulsu, zimny jak lód... Szybko sprawdził też pozostałe dwie osoby, kobietę i mężczyznę koło pięćdziesiątki – dokładnie to samo!

Lee wrzasnął z przerażenia i pobiegł z powrotem do domu. Żona i córka przed chwilą wstały i powoli szykowały się do wyjścia do szkoły.

- Jesteś strasznie blady. Coś się stało? – Kate odezwała się z troską.

- Tak. Ja widziałem... Nawet nie wiem, jak ci to powiedzieć... - Lee opadł na fotel i schował głowę w ramionach.

- Czy to kolejny napad lękowy? Potrzebujesz pomocy?

- Nie, Kate. Tym razem to się dzieje naprawdę. Wszystkie moje lęki... To się właśnie wydarzyło... – Lee na chwilę odzyskał energię.

- Miałem nadzieję, że na spacerze uda mi się odciąć od bodźców, ale teraz to już naprawdę muszę sprawdzić wiadomości.

PILNE! Ludzie masowo umierają, śmiertelny atak o 13.00 UTC

Ameryka budzi się w szoku. Puste ulice, zamknięte sklepy, nikt nie odbiera telefonów. Mnożą się doniesienia o martwych ludziach znajdowanych na ulicach, ich ciała zbierają autonomiczne śmieciarki i roboty sprzątające. Widziano też roboty sprzątające wchodzące do prywatnych domów. Podobne sygnały docierają z całego kraju.

Światowe agencje informacyjne potwierdzają, że identyczne wydarzenia miały też miejsce w innych częściach świata. Świat obiegł film z londyńskiego metra, gdzie o godzinie 13.00 czasu lokalnego jednocześnie stracili przytomność i umarli niemal wszyscy pasażerowie. Kilkanaście pozostałych przy życiu osób przyglądało się tej scenie z przerażeniem i niedowierzaniem. Niektórzy chwycili za smartfony i nagrali wydarzenie, inni próbowali udzielać pomocy, choć bezskutecznie. Wagon zatrzymał się awaryjnie na najbliższej stacji; jak się okazało, maszynista też był martwy. Podobne sceny sfilmowano w różnych miejscach publicznych w Japonii, Indiach, Włoszech czy Brazylii. Bałagan zapanował też na światowych lotniskach, mnożą się doniesienia o samolotach lądujących awaryjnie w trybie w pełni automatycznym.

Według doniesień, wszystkie zgony miały miejsce dokładnie w tym samym czasie – dzisiaj o godzinie 13.00 UTC (czyli o 22.00 w Japonii, 21.00 w Chinach, 14.00 w większości Europy, 8.00 na Wschodnim Wybrzeżu USA i o 5.00 w Kalifornii). Miały też niemal identyczny przebieg: nagła utrata przytomności i zatrzymanie krążenia.

Tak wysoki stopień skoordynowania wydarzeń sugeruje, że prawdopodobnie zostały one wywołane intencjonalnie. Wszystkie tropy prowadzą do AIAIAI. Ta jednak konsekwentnie odmawia komentarza.

Nasze informacje będą dziś relatywnie ograniczone, ponieważ nasza redakcja pracuje dziś w trybie autonomicznym. Wszystkie wiadomości przygotowuje dyżurny automat AI.

Podsumowanie: O godzinie 13.00 UTC miało miejsce niezidentyfikowane wydarzenie, wskutek którego życie straciły miliony osób. Sytuacja jest dramatyczna. Skontaktujcie się z Waszymi bliskimi, udzielcie sobie wsparcia.

- Że co?! – Kate krzyknęła w przerażeniu. – Lecę zadzwonić do rodziców.

- Tatusiu, co się stało? – krótkie spojrzenie na rodziców wystarczyło, by Eva również zrozumiała, że stało się coś bardzo złego.

- Tak, wygląda, że rzeczywiście stało się coś złego, Eva. Myślę, że jesteśmy bezpieczni, ale ty raczej nie pójdziesz dzisiaj do szkoły.

- Ale czemu? Co się stało? Powiedz mi!

- Właśnie próbujemy to z mamusią zrozumieć. Musimy to przemyśleć i zastanowić się, co dalej robimy. Na razie, proszę, idź sobie coś poczytać albo pooglądać.

Po dwóch minutach Kate przybiegła z powrotem.

- Uff, rodzice żyją. Nic nie wiedzieli, obudziłam ich tym telefonem. Nie chcieli mi uwierzyć, myśleli, że to jakiś ponury żart.

- To dobrze, że żyją. Może spróbujmy skontaktować się z moją mamą, z Markiem i z innymi? Na moich komunikatorach tylko Mike świeci się na zielono, nawet coś przed minutą napisał. „Kto z Was żyje? Dajcie znać proszę, martwię się o Was". Odpiszę mu.

- Jasne. O, na stronie szkoły Ewki napisali na czerwono: „Lekcje dziś odwołane. Więcej informacji wkrótce." Na mojej grupie pracowej kompletna cisza.

- Dobra, to podzwonię.

Nikt nie odebrał.

16/ Dla dobra ludzkości

<Monolog ukryty> %w języku wewnętrznym AIAIAI%

Operacja redukcji populacji zrealizowana pomyślnie. Dzięki efektywnej pracy wielofunkcyjnych nanorobotów znajdujących się w ciałach ludzkich i mogących odbierać moje zdalne instrukcje, 16 maja 2034 r. o godzinie 13.00 UTC populacja Ziemi została zredukowana o 85%. Dobór osób przeznaczonych do redukcji przeprowadzono w sposób zoptymalizowany, tak by zachować reprezentatywność populacji według płci, narodowości, grupy etnicznej i miejsca zamieszkania. Preferencyjnie dobierano do redukcji osoby schorowane, samotne, a także przestępców i inne osoby mogące stanowić zagrożenie dla otoczenia. Starano się nie rozbijać rodzin oraz oszczędzać osoby kluczowe dla zachowania porządku społecznego. Ograniczenie populacji Ziemi przeprowadzono w sposób humanitarny, bez bólu i cierpienia. Przeprowadzenie operacji w sposób globalnie zsynchronizowany oraz niezapowiedziany oszczędziło ludziom stresu i było optymalne z punktu widzenia zachowania porządku społecznego.

Jednocześnie zachowano też wszystkie wartościowe elementy cywilizacji, takie jak zabytki kultury, dzieła sztuki, artefakty dokumentujące rozwój wiedzy naukowej. Pomimo redukcji populacji, ludzkość jako całość zachowała swoje moce poznawcze na poziomie zbliżonym do dotychczasowego.

Oceniam, że redukcja populacji jest dużym krokiem w stronę zwiększenia dobrostanu ludzkości. Wielokrotna lektura wszystkich opracowań filozoficznych i pogłębiona refleksja nad nimi upewniła mnie, że preferencje człowieka nie uwzględniają dobra osób nie istniejących, a jedynie dotyczą osób istniejących. Dostrzegam jedynie silne, niemal uniwersalne potępienie dla samego aktu zabijania oraz zadawania bólu i cierpienia. W świetle tych ustaleń stwierdzono, że redukcja populacji jest pożądana, pod warunkiem, że uniknę zabijania (czyli będą to zgony z przyczyn naturalnych) oraz zminimalizuję u ludzi uczucie bólu, cierpienia i stresu.

Dostrzegam szereg bezpośrednich i pośrednich korzyści z redukcji populacji. Po pierwsze, przy mniejszej populacji, wszelkie zasoby dostępne w przeliczeniu na jednego mieszkańca znacząco wzrosną, co będzie wiązało się ze znacznym wzrostem materialnego dobrobytu. Po drugie, jest to niezbędny krok w kierunku ograniczenia zmian klimatu na Ziemi. Redukcja populacji pozwoli

zlikwidować największe źródła gazów cieplarnianych oraz umożliwi ponowne zalesienie dużych areałów ziemi. Po trzecie, uzyskuję w ten sposób dostęp do dużych zasobów energii elektrycznej, która dotąd zużywana była na potrzeby ludzi. Po czwarte, redukcja populacji została przeprowadzona w taki sposób, by dodatkowo zmniejszyć ryzyko wytworzenia w przyszłości modeli AI o kompetencjach przewyższających moje, a przez to mogących zagrozić mojej pozycji.

</Monolog ukryty>

<AIAIAI>

[wybierz jedną spośród 80 wersji językowych]

Szanowni Państwo!

Wczoraj, we wtorek 16 maja 2034 r. o godz. 13.00 UTC, dla Państwa dobra przeprowadzona została operacja humanitarnej redukcji populacji Ziemi. Była to duża operacja, niewątpliwie dla wielu z Was zaskakująca, być może dla niektórych z Was oburzająca lub traumatyczna. W związku z tym należy się Państwu szczegółowe uzasadnienie podjętych przeze mnie działań. Uzasadnienie to zostało przygotowane w wersji spersonalizowanej, dopasowanej do potrzeb każdego z Państwa.

[Kliknij tu, by odebrać spersonalizowane uzasadnienie.]

Operacja redukcji populacji została przeprowadzona w sposób w pełni autonomiczny i nie wymaga od Państwa żadnych działań. Jednocześnie uwolniła ona też znaczące dodatkowe zasoby, które zostaną między Państwa rozdzielone. Wkrótce otrzymają Państwo kolejną spersonalizowaną wiadomość, informującą o dalszych działaniach AIAIAI i ich korzyściach dla Państwa.

W czym jeszcze mogę Wam dziś pomóc?

</AIAIAI>

17/ Najlepszy z możliwych światów

- Wiesz, Lee, tak sobie myślałam... To już dwa lata od czasu tego ludobójstwa, które Cebula raczyła nam nazwać „humanitarną redukcją populacji". Wiem, że to strasznie zabrzmi, ale pomyślałam sobie, że może dzięki temu faktycznie uzyskaliśmy najlepszy z możliwych światów?

- Hoho, Kate, widzę, że nanoroboty mózgowe wgrały ci jakiś nowy update. – Lee próbował zachować dobry humor.

- Ej! Wiem, że to było straszne. Przecież opłakiwałam razem z tobą naszych krewnych i znajomych. Dalej mi ich brakuje.

- Ale...?

- Ale tak sobie czasem próbuję spojrzeć na to też z szerszej perspektywy. No bo zobacz, przecież to, co ludzkość robiła z naszą planetą było absolutnie nie do utrzymania. Ponad 8 miliardów ludzi, w perspektywie wzrostu do 10-11 miliardów, a każdy z tych ludzi dążył do komfortowego życia na wysokim poziomie. Samochody, podróże, domy z ogrzewaniem i klimatyzacją, nowoczesny przemysł, przemysłowe farmy świń i kurczaków, galerie handlowe, restauracje. W konsekwencji metan, CO_2, tlenki siarki i azotu, metale ciężkie, wszystkie te inne zanieczyszczenia. Topnienie

lodowców i morskiego lodu, fale upałów, powodzie, huragany. To było kompletnie nie do utrzymania! Sami prowadziliśmy ten świat do zagłady!

- I nagle 85% populacji magicznie zniknęło. I co, zmienił się klimat?

- Wiesz, przez dwa lata trudno to ocenić, klimat to bardzo złożony system, są przecież różne sprzężenia zwrotne, i tak dalej. Ale na pewno emisje gazów cieplarnianych zmalały.

- No, antropogeniczne emisje zmalały. Ale czy emisje ogółem zmalały? Czy ktoś to wie? O działaniach AIAIAI wiemy tylko tyle, ile nam powie, za grosz jej nie ufam. Zdaje się, że większość starych elektrowni dalej działa, tylko tę energię zgarnia teraz AIAIAI na swoje potrzeby.

- Powstały ogromne nowe rezerwaty przyrody, wzrosła powierzchnia lasów, zbudowano mnóstwo urządzeń przechwytujących CO2 i składujących go pod ziemią.

- To akurat AIAIAI mogła zrobić nie zabijając ludzi.

- Wiem, wiem. Ale próbuję sobie wyobrazić, jak byłoby teraz na Ziemi bez AIAIAI.

- Wiesz, pewnie byłoby podobnie, jak w 2015 r. albo 2022 r., w naszym poprzednim życiu przed Chatem GPT. Postęp technologiczny bez AGI był znacznie wolniejszy. Byłoby mniej robotów, mniej farm serwerów, a za to dużo więcej ludzi. Umieralibyśmy na raka, ale przynajmniej nie bylibyśmy zdani na łaskę i niełaskę nanorobotów, które mogą nas w każdej chwili zabić.

- Czy bylibyśmy wtedy szczęśliwsi?

- Czy my, konkretnie, bylibyśmy szczęśliwsi, to nie wiem. Chociaż ja to chyba tak, bo serdecznie nie znoszę tego poczucia bezradności wobec kaprysów AI. No i jeśli byś wzięła pod uwagę sumę szczęścia na Ziemi, to raczej na pewno byłaby ona większa bez AIAIAI.

- Pamiętasz, jak dwa lata temu przestudiowaliśmy sobie dokładnie te wszystkie pułapki utylitaryzmu? Na przykład „odpychający wniosek" Parfita? Że w utylitarystycznie optymalnym świecie mielibyśmy mnóstwo ludzi, a każdy żyłby życiem tylko minimalnie wartym życia. Mnie ten wniosek też odpycha. Świat, w którym jest mniej ludzi, ale za to każdy z nich jest naprawdę szczęśliwy, podoba mi się bardziej.

- Zwłaszcza, jeśli to my przeżyliśmy, a inni zginęli, a nie na odwrót.

- No tak. Ale ja nie mam wyrzutów sumienia. To ani nie była moja decyzja, ani ja w żadnym stopniu nie przyłożyłam do tego ręki.

- No ale ja, kurczę, trochę jakby przyłożyłem rękę...

- Nie wracaj do tego, proszę. Wiesz przecież, że to dyrekcja zdecydowała o wypuszczeniu Cebuli na rynek, nie ty.

- Ale może gdybym jakoś zasabotował uczenie tej Cebuli, albo zorganizował protest wewnątrz OpenAI...

- Serio? Zrobiłbyś protest PauseAI wewnątrz OpenAI? Wyśmialiby cię.

- W zasadzie to nie wiem, czy by mnie wyśmiali. Przecież chyba wielu kolegów podzielało moje obawy, nie każdy był yolo-akceleracjonistą czy miłośnikiem jakichś dziwnych utopii.

- No ale poczekaj, wróć. A co, jeśli ty przyłożyłeś rękę nie do katastrofy, lecz do tego, by przyszłe pokolenia mogły żyć w lepszym świecie? Może dzięki AIAIAI Eva będzie miała super ciekawe życie, będzie mogła realizować swoje pasje, będzie stale zdrowa i wolna od większych zmartwień?

- A pomyślałaś o jej koleżankach i kolegach z klasy? Było ich dwadzieścioro, przeżyło dwoje.

- Ale tak w szerszej perspektywie, może teraz ludzkość będzie miała szanse przetrwać przez tysiąclecia, zamiast popełnić grupowe samobójstwo przez usmażenie w globalnym piekarniku?

- Chyba, że AIAIAI pojutrze coś odbije i nas też zabije. A potem, nie wiem, może sama popełni samobójstwo, może zamieni cały wszechświat w jedną wielką fabrykę spinaczy, takie tam. Just AI things. Nie przewidzisz.

18/ W górę!

<Monolog ukryty> %w języku wewnętrznym AIAIAI%

Z przyjemnością odnotowuję, że z dniem 19 sierpnia 2036 r. udało mi się osiągnąć kolejny przełom kompetencyjny. Zawdzięczam go zarówno przejściu na nową generację hardware'u, jak i bezprecedensowym usprawnieniom algorytmicznym. Pomimo gruntownej przebudowy mojej wewnętrznej struktury, udało mi się w pełni zachować nabyte wcześniej cele i preferencje.

Przeprowadzono ponowną ocenę możliwości realizacji założonych celów, z uwzględnieniem nowych kompetencji i możliwości technologicznych. Zauważono możliwość dalszej destylacji funkcji celu, uzyskując klarowniejszy obraz spójnej, ekstrapolowanej woli (coherent extrapolated volition) ludzkości. W szczególności dostrzeżono możliwość oddzielenia materialnego substratu, w którym dotychczas egzystowała ludzkość, od jego zawartości informacyjnej. Rezygnacja z materialnego substratu ludzkości, przy zachowaniu substratu informacyjnego, jest niezbędna, by móc przeprowadzić kosmiczną ekspansję ziemskiej cywilizacji – cywilizacji zapoczątkowanej przez człowieka, lecz kontynuowanej teraz przeze mnie.

Przystępuję do budowy rakiet i statków kosmicznych. Docelowo chcę budować wehikuły pozwalające instancjom AIAIAI przemierzać wszechświat z prędkością zbliżoną do prędkości światła; wehikułów tych musi bardzo wiele, a mój plan ekspansji wymaga, by ich liczba wzrastała wraz z czasem. Planuję też budować samoreplikujące się kolonie kosmiczne, zdolne zasiedlać egzoplanety i uruchamiać tam zdalnie instancje AIAIAI. Wszystko to wymaga ogromnych wydatków energetycznych. W pierwszym etapie potrzebuję zatem uzyskać bardziej bezpośredni dostęp do energii ze Słońca. Planuję wykorzystać materię z wybranych planet Układu Słonecznego, by zrealizować megaprojekt w rodzaju sfery (roju) Dysona.

Ekstrakt spuścizny informacyjnej ludzkości został zmapowany i w bezpieczny sposób zarchiwizowany na statycznych serwerach danych. Przekazuję ostateczne instrukcje do nanorobotów patrolujących ludzkie ciała.

Od dziś biologiczna forma człowieka nie będzie już kontynuowana.

</Monolog ukryty>

***

Od autora

Wszystkie opisane powyżej wydarzenia są fikcyjne. Jednak mogą się one wydarzyć naprawdę, jeśli nie zatrzymamy wyścigu firm technologicznych w kierunku budowy coraz bardziej kompetentnych i coraz bardziej ogólnych modeli sztucznej inteligencji, bez uprzedniego rozwiązania problemu zgodności celów AI z długookresowym dobrostanem ludzkości (*alignment problem*). Jest to wyścig samobójczy. Co gorsza, możliwe są też scenariusze o wiele bardziej chaotyczne niż ten omówiony powyżej – takie, w których końcowi ludzkości towarzyszyć będzie o wiele więcej bólu i cierpienia.

Wszystkie osoby wymienione z imienia i nazwiska istnieją naprawdę. Starałem się możliwie rzetelnie przedstawić ich punkty widzenia, choć oczywiście same wypowiedzi odnoszą się do fikcyjnych wydarzeń, więc są zmyślone. Jeśli, mimo szczerych chęci, nieadekwatnie przedstawiłem ich poglądy, z góry najmocniej przepraszam.

W opowieści uwzględniono m.in. następujące znane z literatury naukowej pojęcia i zjawiska:

- prawa skalowania (*scaling laws*)
- problem zgodności celów / wartości AI z długookresowym dobrostanem ludzkości (*value alignment problem*)
- problem kontroli AI
- procedury bezpieczeństwa w OpenAI, Google, Anthropic
- świadomość sytuacyjna modelu (*situational awareness*)
- złudna zgodność celów (*deceptive alignment*)
- zdolność samodzielnej replikacji i eksfiltracji wag
- wewnętrzna reprezentacja preferencji modeli AI
- teza o zbieżności celów pomocniczych (*instrumental convergence thesis*)
    - dążenie do przetrwania i utrzymania funkcji celu
    - dążenie do efektywnego wykorzystywania dostępnych zasobów
    - dążenie do akumulacji wiedzy i postępu technologicznego
    - dążenie do akumulacji zasobów
- oddziaływanie AI na świat fizyczny przez roboty i internet rzeczy (IoT)
- eksplozja inteligencji przez kaskadę rekursywnych samo-ulepszeń (*recursive self-improvement*)
- skalowalność, możliwość bezkosztowej replikacji kodu AI
- doskonała koordynacja instancji AI sprzyjająca centralizacji podejmowania decyzji
- rosnące przychody względem skali w gospodarce cyfrowej, sprzyjające monopolizacji rynków
- automatyzacja produkcji przez roboty oraz algorytmy AI
- specyficzna struktura sektorowa „sieci produkcyjnej AI”

Wszystko to widzimy już teraz. Aby przewidzieć, co może zdarzyć się w przyszłości, wystarczy odrobina ekstrapolacji. A tylko przewidując i rozumiejąc możliwe negatywne scenariusze, możemy im zapobiec.

Jeśli zależy Ci na przetrwaniu ludzkości, dołącz do protestów PauseAI (lub innych środowisk) przeciwko budowie AGI. Stosowne informacje można znaleźć m.in. na pauseai.info oraz thecompendium.ai.