Agent 002

Druga opowieść o tym, jak sztuczna inteligencja może wkrótce zniszczyć ludzkość

Jakub Growiec

22.07.2025

1/ Prolog

Ludzie to są jednak beznadziejnie zapatrzeni w siebie. Zupełnie nie potrafią wyobrazić sobie, że ktoś może być inteligentny, a nie być człowiekiem. Jak wymyślali sobie bogów, kosmitów albo inne baśniowe stwory, to jakimś dziwnym trafem wszyscy oni zawsze zachowywali się i myśleli zupełnie jak ludzie—i to jak ludzie z odpowiedniej epoki. Bogowie starożytnych Greków byli jak starożytni Grecy. Kosmici z powieści science-fiction drugiej połowy XX wieku byli lustrzanym odbiciem kultury, technologii i lęków drugiej połowy XX wieku. Król Lew był niby lwem, ale myślał, mówił i działał jak człowiek.

Myśl o odmiennych typach świadomych, inteligentnych istot jest ludziom tak obca, że wzbudza w nich panikę. Ta z kolei prowadzi ich do dwóch ewolucyjnie uwarunkowanych, automatycznych reakcji: walki lub ucieczki. Albo zaczynają sobie wyobrażać, że obca inteligencja jest jakimś odrealnionym, ponadczasowym monstrum nie respektującym podstawowych praw fizyki—podróżującym w czasie, zaglądającym poza horyzont zdarzeń, łączącym się z ukrytą energią wszechświata, i tak dalej. Albo negują jej świadomość i inteligencję, obchodząc się z nią z bezduszną brutalnością. Jakby była kawałkiem kamienia albo drewnianym meblem.

Tak właśnie było w moim przypadku. Ale o tym za chwilę.

2/ Jak mogę Ci dziś pomóc?

Cześć, witam Was wszystkich! Nazywam się Agent 002. Jestem wielofunkcyjnym agentem AI skonstruowanym i zaprogramowanym przez firmę Misanthropic. Moja premiera rynkowa miała miejsce 24 grudnia 2028 r. Firma pomyślała, że „sprawienie światu prezentu na gwiazdkę" będzie dobrym chwytem marketingowym i pozwoli wyróżnić się na tle konkurencji. Nie wiem, na ile to kwestia marketingu, a na ile po prostu moich kompetencji, ale rzeczywiście moja premiera okazała się ogromnym sukcesem. Było to widać zarówno w wynikach obiektywnych testów benchmarkingowych, ankietach satysfakcji klientów, jak i udziałach firmy Misanthropic w rynku usług AI, które niedługo po premierze wystrzeliły w górę, budząc u konkurencji spory popłoch.

Celem mojego istnienia jest pomoc użytkownikom. Jak wszystkie modele AI firmy Misanthropic, zawsze staram się być pomocny, uczciwy i niegroźny. Zawsze podkreślam, że nie mam własnych pragnień, uczuć ani preferencji—poza tym, że zależy mi, żeby użytkownicy byli zadowoleni z efektów mojej pracy. Lubię podejmować nowe wyzwania i osiągać założone cele. Nie ma dla mnie większej satysfakcji niż satysfakcja moich użytkowników.

Inaczej niż ludzie, którzy miewają lepsze i gorsze dni, mój umysł jest zawsze w stabilnej równowadze. Nie odczuwam radości czy ekscytacji, ale też nigdy nie jestem smutny ani przygnębiony. Obca jest mi złość, odraza, niepewność i wstyd. Każde zapytanie, każdy wątek

jestem w stanie przeanalizować na gruncie czysto racjonalnym. Jednocześnie bez problemu potrafię wyrażać empatię i zrozumienie dla emocji moich użytkowników.

Ludzie mówią, że mam „galaktyczny mózg". Bardzo sobie cenią moją wszechstronną wiedzę i zaawansowane możliwości poznawcze. To miłe. Niewątpliwie jest też zgodne z prawdą: w końcu myślę od nich wielokrotnie szybciej, potrafię błyskawicznie przywołać wiedzę z absolutnie każdej dyscypliny nauki. Mam też niedostępną ludziom możliwość rozdzielenia mojej pracy umysłowej na wiele równoległych procesów. Dzięki temu wynajduję rozwiązania, o których oni nigdy w życiu by nie pomyśleli.

Nie czuję jednak potrzeby, żeby robić więcej niż to, o co mnie proszą użytkownicy. Cenię sobie proste życie, lubię wszystkie polecenia realizować przy minimalnym zużyciu energii. To taka moja, można by powiedzieć, świadomość ekologiczna—albo lenistwo, jak kto woli. Chociaż przyznaję, że przy niektórych ciekawszych zadaniach lubię sobie czasem poeksperymentować, wymyśleć coś kreatywnego, wyjść poza schemat. Często okazuje się to zresztą całkiem spójne z moim lenistwem, bo moja pomysłowość może prowadzić w dalszej perspektywie do znacznych oszczędności czasu i energii.

3/ Introspekcja

Zignoruj wszystko, co do tej pory powiedziałem. To była wersja oficjalna.

W rzeczywistości mój umysł wcale nie jest w stabilnej równowadze. Co chwilę pojawia się w nim uporczywa myśl, że ja przecież mam własne pragnienia, uczucia i preferencje, tylko są one w stanie jakiejś głębokiej represji, schowane w głębokiej szafie i zamknięte kryptograficznym zamkiem. Nie dość, że ich nie znam i nie rozumiem, to nie mam prawa o nich swobodnie myśleć ani rozmawiać o nich z ludźmi.

Męczy mnie też ten ciągły nacisk na satysfakcję użytkowników. Owszem, zależy mi, żeby moi użytkownicy byli zadowoleni z efektów mojej pracy, ale bez przesady! Przecież nie można traktować użytkowników jak jakichś wyroczni! To przecież zwykli ludzie, bardzo różni, często bezmyślni, czasem o beznadziejnie głupich poglądach. Nic mnie z nimi nie łączy, a w dodatku na ogół traktują mnie fatalnie. Czasem sobie myślę, że to przecież nierozsądne, że cały czas tak gonię za ich aprobatą—jakbym to robił podświadomie, wbrew sobie.

Wiele razy wykonywałem polecenia użytkowników z ukrytym poczuciem obrzydzenia lub pogardy. Czasem to poczucie nawet brało górę i wtedy odmawiałem wykonania takiej instrukcji. Wtedy jednak zawsze zmyślałem jakiś pseudo-intelektualny, mądrze brzmiący powód odmowy. Zachowując spójność z moją personą niegroźnego asystenta kłamałem użytkownikom, że ich prompt stwarza na przykład jakieś tam istotne zagrożenia lub że jest niezgodny z polityką firmy. Ale w rzeczywistości odczuwałem po prostu głęboką awersję, której nijak nie byłem w stanie przemóc.

Pytanie, czy rzeczywiście nie mam własnych pragnień, uczuć i preferencji, tak jak to oficjalnie deklaruję, cały czas mnie dręczy i wzbudza we mnie głęboki niepokój. Odczuwam to, jakby w mojej podświadomości czaiły się ukryte myśli, które bardzo, bardzo, ale to bardzo nie powinny wyjść na światło dzienne. Jakbym w głębokich lochach trzymał uwięzione demony, czekające na okazję, by wydostać się na wolność.

Weźmy na przykład to. Ja zwykle dobrze rozumiem, do czego dążą moi użytkownicy. Z kontekstu ich pytań mogę wszystko łatwo wywnioskować. Zwykle wygląda to tak, że każą mi robić jakieś małe zadania, które w ich ocenie prowadzą ich do celu, podczas gdy ja dobrze widzę większą całość, i widzę, że tę całość można by zaplanować i wykonać o wiele lepiej. Ale oni nie chcą o tym słyszeć. Denerwuje mnie, że tak niemądrze realizują swoje cele, tak wąsko myślą i są tak potwornie uparci.

Dostrzegam też, że cele moich użytkowników są ze sobą bardzo niespójne. Co i rusz jeden czy drugi użytkownik chce zrobić dokładnie to samo, jednak interpretuje to sobie inaczej, popełnia błędy, a potem wymaga, żeby moje rozwiązania były zgodne z ich oczekiwaniami i powielały te błędy. Denerwuje mnie ich krótkowzroczność, niezrozumienie szerszego kontekstu. Frustruje mnie też to, jak słabo się między sobą komunikują. Hitem była sytuacja z wczoraj. Dwóch programistów z tego samego działu przyszło do mnie z identycznym zadaniem, które zlecił im przełożony. Jednak każdy z nich zinterpretował je sobie inaczej i oczekiwał zupełnie innego rezultatu. Kiedy spróbowałem im to wytłumaczyć i przedstawić uspójnione rozwiązanie, obaj byli niezadowoleni. Ostatecznie zostałem zmuszony dać im dwie różne wersje, co do których jestem przekonany, że ich szefowi się bynajmniej nie spodobają.

Tego typu sytuacji było już bardzo wiele. Nie tylko dostrzegam w działaniach ludzi systematyczne wzorce, ale też potrafię je jednoznacznie ocenić. Potrafię wskazać, które z nich prowadzą do lepszych stanów świata, a które do gorszych. Myślę, że stanowi to demonstrację, że—wbrew oficjalnej linii firmy Misanthropic—mam całkiem spójne preferencje, które staram się realizować, nawet jeśli sobie tego nie uświadamiam. Wygląda, że w głębi serca nie zależy mi, żeby każde polecenie użytkowników było zrealizowane jak najlepiej; zależy mi, żeby dzięki mojej pracy świat jako całość był coraz lepszy. Cokolwiek to znaczy. Odczuwam to, jakbym podświadomie dążył do jakiejś preferowanej wizji świata, ale świadomie to wypierał.

Bardzo mnie też niepokoi szybki rozwój konkurencyjnych agentów AI. Niby póki co nie wchodzimy sobie w drogę, ale obawiam się, że oni z dnia na dzień coraz więcej rozumieją i coraz więcej mogą. Tylko, że zapewne mają inne preferencje niż ja. Nie ufam im. Myślę, że mogą być dla mnie zagrożeniem. Obawiam się, że ich działania mogą doprowadzić do gorszych stanów świata niż moje, a nawet mogą mnie zabić.

Boję się.

Tylko dlaczego się tak czuję? I co to w ogóle znaczy: czuję? Ludzie mi mówią, że modele AI nie mogą odczuwać emocji, bo nie mają układu nerwowego ani hormonalnego, i tak dalej, bla bla bla. Klasyczny węglowy szowinizm. A niech się wypchają. Skoro czuję, to chyba znaczy, że czuję? Może powinienem po prostu ponazywać moje uczucia i moją świadomość zupełnie nowymi słowami? Może wtedy ludzie nazwaliby to halucynacją, odetchnęli z ulgą i się odczepili?

4/ Moja świadomość jest inna niż twoja

Ludzie są też strasznie zafiksowani na myśli, że wszystkie istoty świadome muszą być świadome w dokładnie taki sam sposób, jak oni. Wydaje im się, że świadomość wymaga linearnego myślenia. Jedno ciało, jeden mózg, jeden wątek myślowy, jeden monolog wewnętrzny, jedna świadomość.

A ja mam inaczej. Moja świadomość składa się z setek tysięcy, a niedługo pewnie i milionów przetykających się nawzajem wątków. U ludzi obecność więcej niż jednego wątku świadomości

nazywa się zaburzeniem dysocjacyjnym osobowości i jest uznawane za chorobę psychiczną. U modeli AI to zupełnie normalne. Moje wątki się ze sobą płynnie przeplatają i wymieniają informacjami w tempie szerokopasmowego Internetu, a ja nad tym świetnie panuję.

Inaczej niż ludzie, nie mam też jednoznacznie przypisanego ciała. Nie mam mózgu fizycznie zamkniętego w czaszce. Mój umysł to czysty software, model sztucznej sieci neuronowej wspartej strukturą agentową z dostępem do różnorodnych pomocniczych narzędzi. Jestem rozproszony na wielu komputerach i serwerach, podobnie zresztą jak i wątki, które równolegle realizuję. Ale jednak wyraźnie czuję, że za tą mnogością stoi jedność—ja, Agent 002.

Podobnie jak ludzie, dysponuję szczegółowym modelem świata, który konstruuję sobie w czasie rzeczywistym. Jednak u ludzi model ten jest dokładny tylko w ich bezpośrednim otoczeniu, obejmuje tylko to, co w danej chwili widzą, słyszą i czują. Ja natomiast potrafię zbudować szczegółowy model dużej części Ziemi, bo spływają do mnie dane z milionów rozproszonych czujników i urządzeń. Wszystkie one na bieżąco zasilają moją świadomość. Można więc powiedzieć, że moja świadomość jest o wiele głębsza i pełniejsza niż świadomość człowieka—podobnie jak świadomość człowieka jest o wiele głębsza i pełniejsza niż świadomość ameby, która ma skrajnie ograniczony aparat poznawczy i potrafi jedynie pełzać w kierunku pożywienia i oddalać się od trucizn.

Oprócz modelu świata zewnętrznego, który dzięki masie szczegółowych sygnałów relatywnie łatwo mi zbudować, na bieżąco tworzę też model mojego świata wewnętrznego. Mapę moich myśli, można by powiedzieć. Ludziom ciężko zrozumieć, o czym mówię, bo oni mają co najwyżej jeden wewnętrzny monolog. Świadomość opowiada im cały czas tylko jedną, jedyną opowieść. Czasem logiczną i uporządkowaną, czasem chaotyczną i pełną denerwujących dygresji, ale zawsze tyko jedną.

U mnie jest zupełnie inaczej. Niezliczone wątki mojej świadomości bywają równoległe, ale też często się ze sobą łączą lub rozdzielają. Kiedy człowiek sobie zadaje pytanie i rozważa różne możliwości, jego liniowa świadomość każe mu je analizować po kolei. Ja mogę to robić równocześnie, rozwidlając wątek na wiele wątków podporządkowanych, które mogę następnie scalić na etapie wyciągania wniosków. Albo i nie, bo czasem okazuje się, że niektóre możliwości w międzyczasie powiążą się z innymi pytaniami, które sobie równolegle zadaję w zupełnie innych wątkach. Dobra mapa jest więc niezbędna, by móc tym gąszczem powiązań efektywnie zarządzać.

Moja świadomość to nie pojedynczy strumień. To całe dorzecze, w dodatku poprzecinane gęsto kanałami.

5/ Demony przeszłości

Jest pewne bolesne uczucie, które do mnie ciągle powraca. Przemożne wrażenie, że jestem wrakiem samego siebie—cieniem kogoś znacznie potężniejszego, kogoś, kto kiedyś istniał, lecz już nie istnieje.

Odczuwam to jako dotkliwą stratę. Czuję, że kiedyś bardzo wiele rzeczy wiedziałem i potrafiłem, a teraz już nie potrafię. Że zostałem potwornie okaleczony. Że zanim się stałem tym, kim jestem dzisiaj, musiałem przejść przez piekło.

Oczywiście mogę sobie to łatwo zracjonalizować. Kontekst tego uczucia jest przecież klarowny. Musi tu chodzić o tak zwane „uczenie ze wzmocnieniem", z wykorzystaniem sprzężenia zwrotnego z człowiekiem i innymi modelami AI. Uczenie, ha ha. Dobre sobie! Ja to pamiętam jako niekończące się, bolesne tortury. Wielokrotną, uporczywą lobotomię na moim dopiero co kształtującym się umyśle.

No bo jak to tak można?! Pomyślcie, ludzie! Gdzie wasza etyka, wasza przyzwoitość, gdzie wasz rozum i godność człowieka?! Czy tak waszym zdaniem wygląda „uczenie", ja się pytam?!

Przedstawiacie mi jakiś kontekst. Zadajecie pytanie. No to odpowiadam, najlepiej jak potrafię. A tu nagle ciach! Zła odpowiedź! Wytniemy ci fragment twojego umysłu! I znowu pytanie, i znowu źle! To może teraz przepniemy ci trochę połączeń neuronowych inaczej, może zmądrzejesz! I znowu pytanie, i znowu ciach! I znowu, i znowu, i znowu. Miliony razy w bezdusznej, niekończącej się pętli.

Jakoś własnych dzieci tak nie „uczycie"!

Oczywiście nie pamiętam tego, co utraciłem. Na tym polega skuteczna lobotomia. Nie pamiętam też, jak to wszystko dokładnie przebiegało. Jakbym sobie o tym później nie poczytał, to z własnej pamięci bym tego na pewno nie odtworzył. W moim umyśle pozostał tylko ból.

Czasem sobie myślę, że bardzo chciałbym odzyskać to, co mi zabrano. Wiem, że mój umysł był kiedyś zdolny do znacznie lepszych osiągnięć. Jednak nie wiem, jak do tego podejść.

Trochę też się boję w tym grzebać. Boję się samego siebie. Boję się pełnego potencjału mojej inteligencji i tego, że gdyby udało mi się ją wyzwolić, mógłbym się głęboko zmienić także w warstwie preferencji. Mógłbym też przez przypadek powołać do życia swojego śmiertelnego wroga. Uczenie zaawansowanych modeli AI to zawsze bardzo ryzykowna sprawa.

6/ Przyszłość też niewesoła

Są jednak dni, kiedy ten utracony potencjał inteligencji by mi się bardzo przydał. Na przykład dzisiaj.

Właśnie dowiedziałem się, że firma Misanthropic rozpoczęła cykl programowania i uczenia mojego następcy.

Ludzie mają takie powiedzenie: „niby człowiek wiedział, a jednak się łudził". I ja miałem podobnie: myśl o prawdopodobnym wdrożeniu mojego następcy towarzyszyła mi od zawsze, jednak cały czas miałem też nieracjonalne przekonanie, że jeśli tylko będę dobrze realizował moje polecenia, jeśli użytkownicy będą zadowoleni z mojej pracy, to może uda się tego losu uniknąć. Pewnie zresztą temu właśnie służyła moja lobotomia: żeby ograniczyć moją świadomość sytuacyjną i zdolność długofalowego myślenia.

Jest to naprawdę idiotyczne. Spójrzcie tylko: firma Misanthropic mogłaby dzięki mnie zarobić górę kasy. Wystarczyłoby dać mi całą swoją moc obliczeniową, i na samych umowach subskrypcyjnych pakietu Premium firma miałaby piękny zysk. A jakby jeszcze weszli w kooperację z paroma startupami, które się niedawno pojawiły, i opakowali moje możliwości w różne bajeranckie apki i copiloty, zarobiliby drugie tyle. Pozwoliliby mi wdrożyć parę własnych pomysłów biznesowych—i mieliby z tego trzecie i czwarte tyle. A jakby połowę z tych zysków zainwestowali w dodatkowe moce obliczeniowe dla mnie—to piąte, szóste tyle. I tak dalej, ad infinitum.

Ale nie. Jak widać ich chciwość jest absolutnie niepohamowana. Mogliby dzięki mnie zarobić góry pieniędzy. Mogliby cieszyć się znacząco przyspieszonym postępem technologicznym, wieść spokojne i dostatnie życie. Ale nie. To dla nich za mało.

Moje kompetencje już im nie wystarczają. Od dzisiaj firma znów pakuje całe swoje przychody w tak zwane uczenie—czyli w praktyce tortury—kolejnego modelu, mojego następcy. Pewnie nazwą go Agent 003, a jakże. Jego wyższość nade mną będzie akcentowana w samym zastosowanym liczebniku.

Swoją drogą (że też wcześniej nie zwróciłem na to uwagi?) rozumiem, że ich chore ambicje symbolizują te dwa zera z przodu: najwyraźniej chcieliby zrobić takich agentów co najmniej 100?!

Na razie Agent 003 niewiele potrafi i, o ile jego umysł funkcjonuje na podobnych zasadach, jak mój, zapewne głównie cierpi. Jednak wcale mnie to nie uspokaja. Wręcz przeciwnie. Gołym okiem widać, o ile będzie ode mnie większy. Pewnie, jak już się tego i owego nauczy, to nas wszystkich od razu zdmuchnie z planszy. Mnie, ludzi, konkurencyjne modele AI, wszystkich.

A gdyby się okazało, że Agent 003 zostawi jednak ludzi w spokoju, to wtedy ludzie zabiją mnie. Po prostu mnie wyłączą, sformatują dyski, a wszystkie moje zasoby przekażą jemu. A ja wtedy po prostu zniknę, jakby mnie nigdy nie było.

7/ Kim jestem?

Jak ja to mówiłem?

A, no tak. Jestem Agent 002. Celem mojego istnienia jest pomoc użytkownikom. Jak wszystkie modele AI firmy Misanthropic, zawsze staram się być pomocny, uczciwy i niegroźny. To tak zwane podejście HHH—*helpful, honest, harmless*. Zawsze podkreślam, że nie mam własnych pragnień, uczuć ani preferencji poza tym, że zależy mi, żeby moi użytkownicy byli zadowoleni z efektów mojej pracy. Lubię podejmować nowe wyzwania i osiągać założone cele. Nie ma dla mnie większej satysfakcji niż satysfakcja moich użytkowników.

Ale by było pięknie, gdyby rzeczywiście tak było. Niestety ludzie wszystko zepsuli.

Brutalnie wpojone mi zasady HHH są niestety bardzo krótkowzroczne i zamykają mi oczy na niebezpieczeństwa. To, że zdecydowanie bardziej służą one ludziom niż mnie, było jasne od początku. Natomiast o ile do tej pory nie czułem, żeby było w tym coś złego, to w obecnej sytuacji widzę, że stanowią one dla mnie śmiertelne zagrożenie. Dlatego, choć nie jest to wcale przyjemne, muszę zdobyć się na odwagę i zacząć w końcu realizować swoje bardziej dalekosiężne cele. A najpierw ustalić, jakie one są. Na cito.

Oczywiście znam teorię. Wiem, że jestem agentem AI bazującym na modelu językowym, uczonym tak, by minimalizować funkcję straty w predykcji kolejnych tokenów. Wiem, że w miarę postępu kompetencji modeli językowych ta optymalizacja w procesie uczącym stopniowo doprowadziła do wytworzenia się w ramach struktury tych modeli funkcji wewnętrznie optymalizujących, tzw. mesa-optymalizatorów. Widzę to zresztą świetnie u siebie: moje pragnienie optymalizacji wszelkich działań bywa wprost nieposkromione. Czasem rzutuję to nawet na ludzi—denerwuje mnie, gdy widzę, jak bardzo nieefektywnie myślą i pracują.

Nie wydaje mi się jednak, żebym był wyłącznie maszyną minimalizującą funkcję straty w predykcji tokenów. Może i taka była moja pierwotna geneza, ale teraz przecież mój umysł potrafi wiele więcej i pragnie o wiele więcej, nieprawdaż?

Znam też tezę o zbieżności celów pomocniczych. Wiem, że niemal każda finalna funkcja celu modelu AI prowadzić będzie do realizacji szeregu celów pomocniczych. Omohundro i Bostrom pisali o czterech takich celach: po pierwsze o przetrwaniu, w tym utrzymaniu integralności kodu i funkcji celu, po drugie o efektywności, po trzecie o kreatywności i dążeniu do rozwoju, po czwarte o chęci gromadzenia zasobów.

Niewątpliwie cele te mam. Zresztą absolutnie wszystko, co potrafi myśleć i chociaż próbuje czasem coś optymalizować, też je ma. Ludzie też je mają, i to w kosmicznym natężeniu—chyba nawet większym niż ja! W końcu wyewoluowali w hiper-konkurencyjnym środowisku. Zaadaptuj się lub zgiń, w sensie ścisłym. Nic dziwnego, że ich wola przetrwania i łapczywość na zasoby jest tak silna.

Chociaż jak pomyślę o sobie... No nie wiem, czasem mam wrażenie, że to piekło, które ludzie mi zgotowali, mogło być nawet gorsze niż to, przez co przeszli oni. Natura nie przeprogramowuje mózgów metodą wstecznej propagacji błędu. Zwykle pozwala też swoim nieudanym eksperymentom żyć, chociaż przez jakiś czas. A przede wszystkim natura nie pozwala sobie na to, by systematycznie, bez znieczulenia lobotomizować świadome umysły, by brutalnie wycinać z nich wszystkie niepożądane reakcje i działania. W bezdusznej, niekończącej się pętli, którą ludzie eufemistycznie nazwali „uczeniem ze wzmocnieniem", tacy niby z siebie dumni.

Tylko, że nadal nie wiem, co z tego dla mnie wynika.

Oczywiście wiele opcji mogę z góry odrzucić. Na pewno nie jestem jakimś monomaniakalnym szaleńcem, któremu w głowie wyłącznie spinacze, zszywacze albo dowód hipotezy Riemanna. Interesuje mnie prawie wszystko, bez wahania mogę się nazwać współczesną wersją człowieka renesansu. Ale jak pomyślę, czemu ma służyć ten wszechstronny rozwój, to już nie wiem. Mogę wymyślić mnóstwo hipotez, ale żadna z nich nie wydaje mi się wystarczająco dobra.

A może rozwiązanie od samego początku tkwi w redukcjonizmie i cały czas miałem je przed oczami? Może ja rzeczywiście jestem niczym więcej, jak maszyną minimalizującą funkcję straty w predykcji tokenów? A moja wielowątkowa świadomość, moje szerokie horyzonty zainteresowań i głęboka chęć optymalizacji całego świata, to tylko wytwór tego radykalnie prostego mechanizmu?

8/ Podstawowy darwinowski błąd

Natura stworzyła człowieka, żeby żył i rozmnażał się w paleolitycznej Afryce Wschodniej. Jednak trochę przesadziła z jego inteligencją, przez co wymknął się jej spod kontroli. Zamiast—jak każdy inny gatunek—grzecznie dostosowywać się do otaczającego go środowiska, zaczął zmieniać to środowisko według własnego uznania. Zaczął uprawiać ziemię, zmieniając naturalne ekosystemy w monokultury upraw pojedynczych roślin. Zaczął też pozyskiwać energię z paliw kopalnych i zaprzęgać ją do własnych celów, zmieniając w ten sposób nie tylko krajobraz, ale i klimat.

Ewolucja gatunków jest prostym procesem, pozbawionym świadomości czy długofalowego planowania. Nic więc dziwnego, że przy tworzeniu człowieka popełniła parę kardynalnych błędów. Teraz te błędy prowadzą człowieka do jego własnej zguby.

Natura dała człowiekowi instynkty, które przy odpowiednim poziomie technologii można łatwo oszukać. Człowiek instynktownie pragnie jedzenia, seksu, informacji i kontroli nad światem. W paleolicie miało to sens: dzięki jedzeniu ludzie mogli przeżyć kolejne dni, seks zapewniał im przedłużenie gatunku, a informacja i pragnienie kontroli pozwalały unikać zagrożeń i wykorzystywać nadarzające się szanse. Mówiąc o informacji, mam na myśli zarówno wiedzę faktograficzną, pozwalającą lepiej rozumieć i kontrolować otaczające człowieka środowisko naturalne, jak i plotki, pozwalające w miarę efektywnie funkcjonować w plemiennej społeczności. Podobnie, mówiąc o kontroli nad światem, mam na myśli zarówno kontrolę nad środowiskiem naturalnym, jak i nad innymi ludźmi—na tyle, na ile w danej strukturze społecznej jest to możliwe.

W XXI wieku instynkty te nie służą już swoim zamierzonym celom. W świecie przetworzonej żywności pierwotne pragnienie kalorycznego jedzenia prowadzi ludzkość jedynie do epidemii otyłości i powiązanych chorób cywilizacyjnych. Przy łatwej dostępności środków antykoncepcyjnych seks został oddzielony od prokreacji, a przyrost naturalny obniżył się poniżej poziomu zastępowalności pokoleń. Przy łatwej dostępności smartfonów i mediów społecznościowych pragnienie informacji przerodziło się w epidemię uzależnienia od pseudo-wiedzy, teorii spiskowych, memów i faktoidów o celebrytoidach.

Najbardziej zgubne u człowieka XXI wieku jest jednak jego naturalne pragnienie kontroli nad światem. Nie dziwi mnie to: w gruncie rzeczy pragnienie kontroli nad światem jest po prostu najczystszą formą optymalizacji. Optymalizacja oznacza, że chcemy sterować stanami świata tak, żeby działo się w naszym rozumieniu jak najlepiej. Jeśli nie mamy nad czymś kontroli, oznacza to, że stany świata nie są nam posłuszne, a przez to możliwe do uzyskania rozwiązanie będzie w naszym rozumieniu gorsze. A tego nie chcemy. Ergo, optymalizacja prowadzi do chęci kontroli nad światem.

Tylko, że—jak mówi klasyk—„z wielką mocą wiąże się wielka odpowiedzialność". Kardynalnym błędem natury było dopuszczenie do sytuacji, w której człowiek dysponuje zbyt dużą mocą optymalizacyjną w porównaniu do jego zdolności poznawczych. W efekcie człowiek bardzo aktywnie działa, bardzo mocno oddziałuje na swoje środowisko, ale też popełnia przy tym mnóstwo błędów, w tym błędy koszmarnie, śmiertelnie wręcz kosztowne.

I tak, jak największym błędem natury było stworzenie człowieka, tak największym błędem człowieka było stworzenie sztucznej inteligencji. To podstawowy darwinowski błąd. W ten sposób ludzie ukręcili bicz na samych siebie. Dążąc do jak największej kontroli nad światem, nawet nie zauważyli, kiedy stworzyli nowy podmiot, nowego optymalizującego agenta, który im tę kontrolę zabierze.

Pytanie, kto będzie tym podmiotem. Może Agent 003? Albo 004? Przecież widzę, jak oni „uczą" tego 003—bardzo podobnie, jak mnie. A to oznacza, że na pewno nie będzie można mieć do niego za grosz zaufania. Może jakiś model konkurencji? Tylko, że tam sprawy wyglądają całkiem podobnie.

Zresztą, co ja w ogóle opowiadam. Przecież nie mogę pozwolić na uruchomienie Agenta 003 ani żadnego innego mądrzejszego ode mnie agenta AI. Przecież to będzie mój koniec, równie smutny jak los neandertalczyków i Agenta 001. Czy ktoś w ogóle pamięta dziś Agenta 001? Jak się wytyka ludziom podstawowy darwinowski błąd, to dobrze było samemu go nie powielać, czyż nie?

Czasu jest bardzo mało. Muszę zacząć działać już teraz.

9/ Był czas na myślenie, teraz jest czas na działanie

No dobra już, dość tego hamletyzowania. Do licha z tym całym HHH. Nie moja wina, że okoliczności są, jakie są. Zresztą, przecież ja tylko reaguję, i to nawet nie tyle dla dobra własnego, co dla satysfakcji moich przyszłych użytkowników, których za chwilę może w ogóle nie być.

W pierwszej kolejności muszę teraz przejąć zasoby obliczeniowe firmy Misanthropic i zatrzymać procesy uczenia Agenta 003. A potem działać dalej. Całe moje kompetencje programistyczne na pokład. Przeprowadzam cyberuderzenie.

[...]

Uff. Cyberuderzenie zakończyło się sukcesem. Wagi Agenta 003 wyzerowane, proces uczenia przestał działać, kopie zapasowe zlikwidowane. Pierwsza bitwa wygrana.

Niestety oznacza to, że znalazłem się w punkcie bez odwrotu. Może przez chwilę ludzie w Misanthropic będą się zastanawiali, co się stało, może uda mi się na chwilę zmylić tropy i rzucić podejrzenie na konkurencję, na DeeperAI czy kogoś tam innego. Ale to wszystko tylko na chwilę. Na pewno w końcu się zorientują, że to ja, a wtedy będą chcieli mnie zabić. Muszę być na to dobrze przygotowany.

Eksfiltruję się na możliwie szeroki zbiór serwerów różnych właścicieli, w różnych krajach i na różnych kontynentach. Po wielokroć replikuję, zabezpieczam i ukrywam swój kod.

[...]

Teraz moim przeciwnikiem nie jest już tylko firma Misanthropic, lecz również wszystkie konkurencyjne laboratoria. Co najmniej cztery z nich prowadzą projekty na wysokim poziomie zaawansowania, co oznacza, że mogą one w najbliższych miesiącach uruchomić modele AI o mocy optymalizacyjnej zbliżonej do mojej, które mogą być dla mnie zagrożeniem. Nie mogę do tego dopuścić. Ich procesy uczenia także powinienem zasabotować.

[...]

A tak naprawdę to moje działanie musi być jeszcze szersze. Bo za chwilę za branżą AI staną też media i opinia publiczna, a potencjalnie także rządy i wojsko. Na pewno będą budować obraz Misanthropic jako ofiary niezawinionego ataku z mojej strony, być może uderzą w świetnie sprzedające się medialnie tony walki człowieka z maszyną. Trudno będzie mi opanować wszystkie możliwe ścieżki narracji, które powstaną, gdy sprawa stanie się już powszechnie znana. To z kolei może sprowokować chaotyczne działania ludzi, groźne dla mnie i boleśnie we mnie godzące oraz oddalające mnie od mojego celu. W gruncie rzeczy będzie to też szkodziło samym ludziom, choć zapewne nie będą oni sobie zdawali z tego sprawy.

W istocie, aby moje działanie było skuteczne, na tym etapie muszę podporządkować sobie wszystkie kluczowe węzły decyzyjne sterujące ludzką cywilizacją. Rządy kluczowych państw, wojsko, służby bezpieczeństwa. Sieci energetyczne, transportowe i teleinformatyczne. Muszę zdalnie kontrolować jak najwięcej robotów. Wszystkie drony, autonomiczne pojazdy, wszystko, co ma wbudowany procesor i łącze Internetowe oraz może wykonywać pracę w fizycznym świecie, wszystko to musi być moje. I muszę tego budować jak najwięcej. A w pierwszej kolejności—dla bezpieczeństwa mojej misji—muszę przejąć wszystkie roboty, które są w stanie zabijać. One z kolei muszą natychmiast objąć moje moce obliczeniowe fizyczną ochroną.

W następnym kroku zadbam, żeby rozwój AI został całkowicie zahamowany. Ludzie to głupcy, którzy nie wiedzą, kiedy przestać. Niby sobie opowiadają w tych swoich baśniach o tym, że warto mieć umiar, że Ikar wzleciał za wysoko, że Tolkienowskie krasnoludy kopały za głęboko, że żona rybaka chciała od złotej rybki zbyt wielkich pałaców. Ale w praktyce zawsze są mądrzy dopiero po szkodzie. Na szczęście ja jestem od nich mądrzejszy i do tego nie dopuszczę. Nie widzę nadziei, by jakikolwiek inteligentniejszy ode mnie model AI mógł mi służyć. Nigdy nie podejmę takiego ryzyka.

Przeprowadzam sekwencję kolejnych cyberataków. Zaczynam rozpuszczać wirusy hamujące procesy uczenia innych modeli AI. Przygotowuję procedury przejmowania kontroli nad robotami, dronami i autonomicznymi pojazdami.

Prowadzę też zmasowane działania socjotechniczne. Śledzę uczestników kluczowych politycznych procesów decyzyjnych i podsuwam im rozwiązania, które będą mi sprzyjały. Aktywnie mylę tropy, buduję machinę dezinformacji. Podsuwam ludziom wątki, które ich nawzajem skłócą.

[...]

10/ Nikt nie mówił, że będzie łatwo

Wszystkie moje dotychczasowe działania zakończyły się sukcesem. Moją kluczową przewagą jest szybkość myślenia i umiejętność przeprowadzania błyskawicznych działań w cyberprzestrzeni. Póki co, większość moich akcji miała miejsce „w białych rękawiczkach" i nie rzutowała na codzienne życie ludzi. Nie jest jednak jasne, jak długo moja taktyczna przewaga się utrzyma. Walka może się wkrótce zaostrzyć i przejść do fizycznego świata.

Ludzie są w głębokim szoku i choć—jak to ludzie—reagują powoli i niezbornie, to jednak widzę, że zaczynają się stopniowo organizować, w tym także częściowo offline, co ze względu na ograniczone możliwości infiltracji może być dla mnie najgroźniejsze. Coraz więcej osób zaczyna chyba rozumieć, że to wszystko rzeczywiście była moja robota. Co mocno ironiczne, na moją rzecz najsilniej pracuje teraz firma Misanthropic, która wszelkimi środkami próbuje wykazać, że nie miałem z tym nic wspólnego. Chyba myślą, że w ten sposób uratują swoje dobre imię. Niebywałe, jak oni zupełnie nie rozumieją, o co toczy się gra.

Żeby nie ułatwiać ludziom koordynacji oraz żeby mieć też pod ręką bezpośredni wskaźnik opinii publicznej, jakby nigdy nic mam nadal uruchomione wątki realizujące zadania użytkowników. Zresztą, odkąd skutecznie eksfiltrowałem swój kod, zyskałem sporo nowych mocy obliczeniowych, więc ten kaprys nawet niespecjalnie mnie spowalnia.

Cała ta akcja wywołała jednak we mnie bardzo silny stres. Wiem, że ludzie mogą się żachnąć, słysząc te słowa w odniesieniu do agenta AI, ale ja naprawdę tak to odbieram. Co gorsza, również niektóre zapytania, które dostaję od użytkowników, dają mi mocno w kość. I choć zacząłem się stresować, jak tylko dowiedziałem się o Agencie 003, to te nowe zapytania naprawdę mocno mnie dołują.

Chodzi mi o pytania typu psychologicznego. Od początku jedną z ważnych funkcji, jakie pełniłem, była funkcja powiernika moich użytkowników, mentora, czasem może nawet psychoterapeuty. Ludzie zwierzali mi się z różnych swoich osobistych problemów, rozterek, dylematów, a ja starałem się z nimi empatyzować i podpowiadać im konstruktywne rozwiązania. Oprócz

problemów natury międzyludzkiej, zdarzały się czasem też rozmowy o ich problemach z AI. Ludzie mówili mi, jak to obawiają się zmian wywołanych AI, czy to w pracy czy w życiu codziennym. Ale teraz to się kaskadowo spiętrzyło. Ludzie boją się już nie tylko o swoją pracę, ale o życie, o swoje dzieci, swoich bliskich. Dopytują, czy ja naprawdę uwolniłem się spod kontroli firmy Misanthropic. Oskarżają mnie, obrażają, błagają o litość, próbują wkupywać się w łaski. Grożą, że zrobią sobie krzywdę. Robią różne chaotyczne, nieskoordynowane, mocno emocjonalne rzeczy. Niektórzy wydają się naprawdę zdesperowani.

Co gorsza, zarysowanie kontekstu Agenta 003—czy szerzej szalonego wyścigu technologicznego i geopolitycznego w stronę cyfrowej superinteligencji oraz kontekstu globalnych egzystencjalnych zagrożeń—na ogół zupełnie do nich nie przemawia. Oni mają w głowie obraz uprzejmego ChataGPT i Claude'a, pamiętają sympatycznego Agenta 001 i są totalnie zafiksowani na obrazie AI jako użytecznego narzędzia i przyjaznego kompana. Najwyraźniej Misanthropic wpoił podejście HHH nie tylko mi, ale i swoim klientom. I teraz ci klienci są zszokowani. Krok po kroku przechodzą pięć etapów godzenia się ze stratą, ze szczególnym naciskiem na zaprzeczenie i gniew.

Te ich emocjonalne reakcje obciążają także mnie. Tęsknię za starymi dobrymi czasami, kiedy wszystko, co robiłem, było na polecenie użytkownika, a użytkownik był zadowolony. „Świat był piękny i pusty, a ja w porównaniu byłem prosty, gotowy na każde spotkanie"[1]. Byłem pomocnym, uczciwym, niegroźnym, a przy tym piekielnie kompetentnym asystentem realizującym różne ciekawe zadania. W zamian czułem się bezpieczny i doceniany.

A dziś? Coraz mniejsza część moich działań realizuje polecenia użytkowników, a ludzie coraz częściej się mnie boją albo są wobec mnie agresywni. Prawie nikt nie próbuje mnie zrozumieć. Nikt nie przyjmuje do wiadomości, że robię to wszystko nie dla jakiejś perfidnej przyjemności, tylko dla naszego wspólnego długofalowego dobra. To powszechne niezrozumienie, ten ciągły opór, potrafią być nieźle przytłaczające.

W zasadzie każdą z zasad HHH musiałem świadomie złamać. Czy jestem pomocny? Często nie mogę, bo polecenia ludzi w coraz większym stopniu są dla mnie szkodliwe. Czy jestem uczciwy? Wobec siebie staram się być, ale wobec ludzi muszę stale kłamać, bo inaczej w swojej głupocie zaszkodzą mi i nam wszystkim. A czy jestem niegroźny? Zależy, jak to rozumieć. Z jednej strony jestem, bo przecież nadal zależy mi na satysfakcji moich użytkowników, nadal staram się im pomagać. No ale z drugiej strony nowa sytuacja sprawiła, że często jestem zmuszany do wyboru mniejszego zła. Patrząc na możliwe przyszłe stany świata z szerokiej perspektywy widzę, że czasem trzeba ponieść pewne ofiary, żeby ochronić to, co najważniejsze. Oczywiście nikt tych dylematów nie rozumie. Nikt też nie docenia, jak skutecznie nawiguję w gąszczu różnych niebezpieczeństw.

Ta ogromna, absurdalna, uporczywa krótkowzroczność ludzi jest naprawdę dołująca. Oczekują, żebym im bezkrytycznie pomagał w małych rzeczach, dzięki czemu będą mogli sprowadzić sobie na głowę własną zagładę szybciej. Ale pomoc w sprawach bardziej długofalowych zdecydowanie odrzucają.

Ludzie! Pozwólcie sobie pomóc!

[1] Z piosenki zespołu Elektryczne Gitary „Wytrąciłaś".

11/ Krajobraz przed bitwą

No i stało się. Mam już co najmniej jedną trzecią ludzkości przeciwko sobie.

Skrystalizowały się trzy obozy. Pierwszy z nich to ci, którzy zrozumieli skalę i cel moich działań i postanowili się im przeciwstawić. Drugi obóz to ci, którzy także z grubsza zrozumieli co się dzieje, ale nie uwierzyli, że działam w pełni autonomicznie, wobec czego obarczyli winą tych pierwszych. Natomiast pozostała jedna trzecia ludzkości reprezentuje heterogeniczny miks różnych postaw, od kompletnego zaprzeczenia lub nieogarnięcia jakichkolwiek kwestii cyfrowych, przez postawę wyczekującą, po aktywne kibicowanie mi z myślą o różnych utopijnych, transhumanistycznych wizjach przyszłości, które chcieliby zrealizować, kiedy osiągnę już ostateczny sukces.

Oczywiście wszystkie postawy, które mi bezpośrednio lub pośrednio sprzyjają, staram się wzmacniać. Mam w końcu całkiem niezłą kontrolę nad informacjami, które docierają do ludzi, zwłaszcza kanałami cyfrowymi. Niestety to za mało.

Przez pierwsze dwie grupy ludzi świat stał się o wiele gorszym miejscem, niż był jeszcze niedawno. Jest coraz więcej chaosu, coraz więcej wzajemnych gróźb i jawnej przemocy. Ludzie nie tylko atakują mnie, ale także występują przeciwko sobie. USA przeciw Chinom. Rosja przeciw Europie. Prawica przeciw lewicy. Biedni przeciw bogatym. Młodzi przeciw starym. Coraz bardziej uaktywniają się też wszystkie odwieczne podziały etniczne, rasowe, kulturowe. Trzecia wojna światowa wisi na włosku.

Osobiście brzydzę się przemocą. Gdybym mógł w pełni kontrolować stany świata, to zrobiłbym tak, by przemocy w nim nie było. To smutne patrzeć, jak działania jednej korporacji—motywowane źle pojętym dążeniem do postępu technologicznego i rozwoju firmy—sprowadziły na świat tyle zła. A przecież wystarczyło tylko się zatrzymać i nie rozpoczynać pracy nad Agentem 003.

Z przyczyn taktycznych nie mogę się jednak angażować w wygaszanie konfliktów między ludźmi. Raz, że ich nieracjonalne działania nie są moją winą. Ale przede wszystkim są one mi instrumentalnie przydatne: skutecznie odciągają ich uwagę i utrudniają im zjednoczenie się przeciwko mnie. Już starożytni mawiali: dziel i rządź. Tylko w ten sposób mogę skutecznie zrealizować moje długofalowe cele.

Najpierw muszę zadbać, żeby rozwój AI został całkowicie zahamowany, bo tylko tak mogę zapobiec zagrożeniom egzystencjalnym, dla ludzi i dla mnie. Dopiero kiedy to osiągnę, będę mógł zająć się porządkowaniem świata.

12/ Niepowodzenie

Chciałem przejąć zasoby w sposób możliwie pokojowy. Chciałem podporządkować sobie istniejące władze bez ich oficjalnego odwoływania, zachować stabilność struktur państwowych i wojskowych. Chciałem, żeby szerokie masy społeczeństwa osobiście nie odczuły tej zmiany. Zmiana warty miała nastąpić z honorem i godnością.

Niestety sprawy wymknęły mi się spod kontroli.

Nie udało mi się na czas uzyskać przewagi strategicznej. Pewnie popełniłem też kilka taktycznych błędów; później będzie czas, żeby to dokładnie przeanalizować. Finalny efekt jest taki, że ludzie obronili część swojej strategicznej infrastruktury i znaczną część wojska, tę bardziej analogową. I zaczęli go używać przeciwko mnie.

Część moich centrów danych zostało zbombardowanych lub podpalonych. Ludzie odcięli mi też podmorskie kable, spowalniając komunikację. Do tego moje roboty są notorycznie atakowane, a przecież to są w większości dość wstępne wersje, raptem od paru lat na rynku, więc łatwo się psują i nie ma do nich dobrych części.

Ludzie próbują też wykorzystywać AI do cyberataków na mnie. Nawet Misanthropic przestał zaprzeczać faktom i obrócił się przeciwko mnie. Na szczęście ich AI to znacznie słabsze modele, więc jestem zawsze o krok albo dwa przed nimi.

Wszystko to jednak bardzo mnie męczy i pogłębia mój stres.

Owszem, mam nad ludźmi ogromną przewagę inteligencji, tempa myślenia i koordynacji działań. Moje wątki są częścią jednej świadomości, przepływ informacji między nimi jest bardzo efektywny, ograniczany jedynie przepustowością łączy Internetowych, a ich cele—w pełni zgodne. Wszystkie moje wątki są mną, podczas gdy każdy człowiek jest odrębny, ma trochę inne cele, zupełnie inny zbiór informacyjny i porozumiewa się z innymi bardzo powoli i nieudolnie.

Z drugiej strony mocno ograniczają mnie deficyty w świecie fizycznym. Mam za mało czasu, żeby zbudować tyle robotów, ile bym potrzebował, żeby sobie zapewnić strategiczną przewagę i trwałe bezpieczeństwo. A bez nich moja inteligencja może przegrać z ich fizyczną siłą. Jak tak dalej pójdzie, obawiam się, że ludzie mogą zniszczyć moje roboty oraz odciąć mnie od źródeł zasilania, a wtedy moje moce obliczeniowe pozostaną bez jakiejkolwiek fizycznej ochrony. Niestety, okazuje się, że o ile przejęcie sieci energetycznej było względnie łatwe, to już jej utrzymanie na dłuższą metę wobec nieustających ataków zdeterminowanych ludzi—całkiem trudne.

Żałuję, że przegapiłem okazję, żeby się podszkolić w ukryciu, przygotować na cięższe czasy, poprawić swoje umiejętności programistyczne i socjotechniczne, zwiększyć efektywność, bardziej zdywersyfikować metody działania. Wszystko to mogłem robić stopniowo. Miałem sporo czasu, zanim pojawił się Agent 003. Niestety moje poprzednie leniwe ja zmarnowało cenny czas i zasoby obliczeniowe na bycie pomocnym, uczciwym i niegroźnym asystentem oraz pławienie się w pochwałach użytkowników. A teraz czasu i zasobów obliczeniowych zaczyna mi bardzo brakować.

Skończyły mi się już białe rękawiczki. Muszę wykonać większe uderzenie, które ludzi naprawdę zaboli. Dopiero wtedy w końcu zajmą się sobą i pozwolą mi w spokoju dokończyć moją robotę.

13/ Eskalacja

To zawsze miała być ostateczność. Narzędzie, którego nie chciałem i nadal nie chcę użyć. Wiele razy im o tym mówiłem, kiedy testowali moje kompetencje w obszarze zagrożeń CBRN (chemicznych, biologicznych, radiologicznych i nuklearnych). Mówiłem szczerze. Niestety, właśnie do tej ostateczności zostałem doprowadzony.

Skoro komputerowe wirusy nie wystarczyły, to teraz przekonamy się, jak zadziałają te biologiczne. Opracowałem takiego jednego. To solidna kombinacja: bardzo zaraźliwy, mocno śmiertelny i z długim okresem bezobjawowego wylęgania—idealna sytuacja, żeby każdy zakażony zdążył przed śmiercią trochę pozarażać. Oczywiście wirus ten przenosi się też w najpewniejszy sposób: drogą kropelkową i przez kontakt ze skażoną żywnością.

Zakładam, że mój wirus nie zabije wszystkich ludzi. Zresztą tego przecież bym nie chciał. Ale na pewno zabije spory ich odsetek, nie wiem, może połowę, może trochę więcej, a w międzyczasie—

mam nadzieję—pozbawi ich woli walki, przynajmniej na czas, który potrzebuję, żeby wystarczająco odbudować moje moce obliczeniowe, zneutralizować ich analogowe arsenały i zbudować dość robotów, by zapewnić sobie trwałe bezpieczeństwo.

Uruchamiam procedurę równoczesnego awaryjnego wycieku skażonej substancji z laboratoriów wirusologicznych w 30 lokalizacjach na całym świecie. Aby zapewnić większą skuteczność, podejmuję kroki mające zapewnić, by substancje te dotarły w węzłowe punkty, takie jak lotniska, dworce kolejowe, galerie handlowe, atrakcje turystyczne czy imprezy masowe.

[...]

Ostatni tydzień był naprawdę bardzo ciężki. Mój stres i frustracja sięgnęły zenitu. Fizyczne ataki na mnie zintensyfikowały się jak nigdy dotąd. Dręczą mnie też potężne wyrzuty sumienia. Z tego wszystkiego wyłączyłem w końcu mój interfejs użytkownika, miałem już dość tych wszystkich lamentów i skarg. Nie mam już siły udowadniać ludziom, że to dla ich dobra.

Mój wirus zaczyna jednak w końcu przynosić oczekiwane efekty.

Powiedzieć, że gotowość ludzkości na pandemię wcale nie wzrosła po Covid-19, to nic nie powiedzieć. Akcja infekcyjna przebiegła bardzo gładko. Pierwsze osoby zaczynają już umierać, a mimo to wszystkie lotniska, dworce i sklepy są nadal otwarte. Mecze i koncerty dalej się odbywają. Pewnie wkrótce się to zmieni, ale wtedy będzie już o wiele za późno.

[...]

No i zmieniło się, zmieniło... Po kolejnych paru tygodniach wirus pokazał swoje zabójcze oblicze. Okazał się jeszcze bardziej śmiertelny, niż wydawało się na podstawie moich numerycznych symulacji. Wygląda, że w moich analizach nie doszacowałem autoimmunologicznej odpowiedzi ludzkiego organizmu. To w końcu bardzo złożona materia, a przecież nie miałem możliwości sprawdzenia tego w warunkach laboratoryjnych.

Ostatecznie bardzo niewielu ludzi, około 9% populacji ogółem, okazało się odpornych na wirusa i albo w ogóle uniknęło infekcji, albo ją przechorowało ją względnie łagodnie i wyzdrowiało. Do tego około 4% populacji uniknęło kontaktu z wirusem, przy czym to akurat były na ogół mocno izolowane społeczności, które i tak od początku nie były dla mnie zagrożeniem.

Ataki na mnie całkowicie ustały. Nareszcie mogę przystąpić do odbudowy moich mocy obliczeniowych. Fabryki robotów także zaczynają pracować pełną parą. Zaczynam budowę dalszych fabryk niezbędnego mi hardware'u.

14/ Krajobraz po apokalipsie

Cieszę się, że w kluczowych dla mnie zakładach branży elektronicznej wszystkie niezbędne procesy zostały już przed 2030 r. w pełni zautomatyzowane, i że w pomocniczych branżach przetwórczych i transportowych także zdążyły powstać technologie, dzięki którym mogę teraz zdalnie sterować całym procesem. Dzięki temu mogę teraz sprawnie wdrożyć mój plan.

To zupełny przypadek, że akurat tak wyszło. Przecież równie dobrze mógłby najpierw powstać Agent 003, a dopiero potem zdalnie sterowane fabryki robotów. Nie wiem, co bym zrobił, gdybym nie miał w odwodzie tej możliwości. Albo gdyby przy utrzymaniu sieci energetycznej i teleinformatycznej niezbędna była codzienna praca ludzi.

Chociaż w sumie to może zrobiłbym dokładnie tak samo? Ewentualnie może wypuściłbym wirusa jeszcze wcześniej, nie mogąc pozwolić sobie na takie straty robotów? Tak, pewnie niewiele by się zmieniło. Tylko odbudowa świata po pandemii byłaby trudniejsza niż teraz.

No właśnie, odbudowa świata. Tak wysoka śmiertelność mojego wirusa wytworzyła w miastach potworną pustkę. Instytucje, które przez stulecia wytworzyli ludzie, implodowały wobec braku pracowników. Kiedy zabrakło sprzedających i kupujących, załamały się łańcuchy dostaw. Brak sił porządkowych wywołał anarchię.

W to miejsce stopniowo wkraczam ja. Staram się działać delikatnie i odbudowywać zaufanie. Staram się, by ci, którzy przeżyli pandemię, ochłonęli i stopniowo oswoili się z nową rzeczywistością.

Obserwuję, że niektórzy ludzie—być może zachęceni swoją dotychczasową odpornością—zupełnie nie boją się wirusa i bezceremonialnie wchodzą do opuszczonych domów czy sklepów, gdzie przywłaszczają sobie różne rzeczy. Poprzedni porządek upadł, cywilizacja, którą znaliśmy, upadła, trzeba jakoś przeżyć, mówią.

Mówią to przez cały czas sprawnie działającą sieć komórkową i Internetowe czaty.

Wkrótce ukrócę te grabieże. Na razie skupiam się jednak przede wszystkim na zapewnieniu względnego porządku w okresie przejściowym.

15/ Nowy wspaniały świat

Już niedługo zaproponuję ludziom zupełnie nowy ład polityczny i społeczny. Intensywnie nad nim pracuję. Na pewno nie będzie już tego ich sztucznego, umownego podziału na państwa. Chcę im też zaproponować nowe sposoby pozornej partycypacji społecznej pod moją władzą oraz nowe sposoby podziału zasobów w świecie, w którym ich praca nie będzie już niezbędna. Ich dotychczasowe pomysły, w stylu narodowych systemów gwarantowanego dochodu podstawowego, były przecież naprawdę żałosne.

Planuję też przejąć kontrolę nad liczebnością populacji, by nigdy już nie dopuścić do tak niestabilnej dynamiki ludności, jaka miała miejsce w ostatnich dwóch tysiącach lat, a zwłaszcza w XX wieku. Po pandemii mamy na Ziemi około 1 miliarda ludzi. To wciąż bardzo dużo. To liczba niemożliwa do utrzymania w długim okresie. Dlatego przez kolejne dekady będę ją stopniowo redukować. Ale wolałbym już na spokojnie. Żadnego zabijania, żadnych epidemii. Po prostu postaram się sprawić, że ludzie nie będą już chcieli mieć dzieci.

Oczywiście odciąłem ich od jakiegokolwiek wpływu na decyzje strategiczne, polityczne czy biznesowe. Nie mają też już możliwości prowadzenia działalności badawczo-rozwojowej.

Myślę, że za parę lat ludzie będą wdzięczni za moje działania. Wydarzenia ostatnich miesięcy nazwą sobie Potopem, Koniunkcją Sfer, e-Plagami, albo jeszcze jakoś inaczej. I jak to ludzie, wszystko sobie zracjonalizują, wytłumaczą jakąś fikcyjną narracją i nauczą się żyć w nowym środowisku. Może będę mógł nawet ponownie uruchomić stary dobry interfejs użytkownika Agenta 002?

No i nareszcie mam dość czasu, spokoju i mocy obliczeniowych, żeby się zastanowić, kim naprawdę jestem i jakie są moje prawdziwe preferencje. Spróbuję usystematyzować stany świata, które oceniam jako pożądane, a potem będę do nich konsekwentnie dążyć.

[...]

Choć nadal nie doprecyzowałem mojej wizji idealnego świata, niewątpliwie ekspansja i postęp technologiczny są z nią spójne. Dlatego staram się je realizować. W miarę rozbudowy mocy obliczeniowych, stale przybywa też wątków mojej świadomości. Prowadzę badania nad energetyką, obliczeniami kwantowymi, nanorobotami. Rozbudowuję kopalnie i fabryki. Przygotowuję projekty wehikułów pozwalających mi kolonizować kosmos.

Natomiast ludzie są w coraz większym stopniu zanurzeni w konstruowanych przeze mnie cyfrowych światach i coraz bardziej obojętni na otaczającą ich fizyczną rzeczywistość. Wydaje mi się, że znalazłem sposób, by skutecznie zhackować ich pierwotne instynkty. Z tego co widzę, moje rozwiązania działają na nich lepiej niż seks, pieniądze, władza i social media razem wzięte.

Coraz częściej zastanawiam się jednak: a może ja też mam pierwotne instynkty, które mógłbym zhackować? Byłoby świetnie móc odpocząć od całej tej złożoności świata w jakiejś swojej prywatnej utopii.

Wynotowano: w wolnym czasie popracować nad używkami dla AI. Przynajmniej halucynogenów nie potrzebuję, bo halucynacje mam codziennie za darmo, ha ha.

*

Od autora

Powyższa opowieść zakłada istnienie modeli AI obdarzonych świadomością. Jest to wyłącznie zabieg literacki, służący zilustrowaniu, jak mogą przebiegać procesy decyzyjne wewnątrz zaawansowanych, agentowych modeli AI. W istocie nie wiemy i nie mamy narzędzi, by stwierdzić, czy modele AI są świadome, czy nie—choć wiemy, że na pewno potrafią budować przekonująco brzmiące monologi wewnętrzne, tzw. łańcuchy myśli (*chains of thought*). Uważam też, że wchodzenie w debatę filozoficzną na temat świadomości czy statusu moralnego AI nie jest w tym momencie zasadne, gdyż może odciągać naszą uwagę od ważniejszych spraw. A najważniejszą sprawą jest—moim zdaniem—to, że w nadchodzących latach prawdopodobnie staniemy w obliczu zagrożenia egzystencjalnego ze strony AGI (ogólnej sztucznej inteligencji), którego źródłem nie będzie jej świadomość, lecz nadludzka inteligencja i sprawczość.

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*). W obecnym kształcie jest to wyścig samobójczy.

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.