Wprowadzenie

Wizyta w centrum projektowania umysłu

Jest rok 2045. Wybierasz się na zakupy. Twoim pierwszym celem jest Centrum Projektowania Umysłu. Po tym, jak przekraczasz próg, pojawia się przed tobą bogate menu, w którym znajdujesz ulepszenia mózgu opatrzone wymyślnymi nazwami. „Umysł roju" to implant oferujący dostęp do najskrytszych myśli twoich najbliższych. „Ogród zen" to mikrochip umożliwiający wejście w stany medytacyjne dostępne wcześniej tylko mistrzom zen. „Ludzki kalkulator" zapewni zdolności matematyczne na poziomie sawanta. Na co się zdecydujesz? Na udoskonalenie uwagi? Zdolności muzyczne na poziomie Mozarta? Możesz wybrać jedno ulepszenie albo zestaw kilku.

Później wstępujesz do sklepu z androidami. Przyszedł czas na kupienie tego nowego androida wykonującego prace domowe. Menu umysłów SI jest rozległe i zróżnicowane. Niektóre z nich mają zwiększone zdolności percepcyjne albo zmysły,

których brak ludziom, inne wyposażone zostały w bazy danych obejmujące cały internet. Starannie wybierasz opcje, które przydadzą się twojej rodzinie. Dzisiaj jest dzień decyzji dotyczących projektowania umysłu.

Niniejsza książka poświęcona jest przyszłości umysłu. Dotyczy tego, jak nasze rozumienie samych siebie, naszych umysłów i naszej natury może radykalnie zmienić przyszłość – na lepsze lub na gorsze. Nasze mózgi wyewoluowały w określonym środowisku i są poważnie ograniczone przez anatomię i ewolucję. Sztuczna inteligencja (SI) otwiera jednak przed nami ogromną przestrzeń projektowania, dostarczając nowych materiałów i sposobów działania, a także nowych dróg eksplorowania jej w tempie znacznie szybszym niż tempo biologicznej ewolucji. Nazywam to ekscytujące nowe przedsięwzięcie *projektowaniem umysłu*. Jest ono formą „inteligentnego projektu", jednak projektantami jesteśmy tu my, ludzie, a nie Bóg.

Perspektywa projektowania umysłu wzbudza we mnie pokorę, ponieważ szczerze mówiąc, ewolucja nie zaprowadziła nas zbyt daleko. Obcy w filmie *Kontakt*, nakręconym na podstawie powieści Carla Sagana, spotykając po raz pierwszy człowieka, mówi: „jesteście bardzo interesującą rasą. Jesteście zdolni do pięknych marzeń i miewacie okropne koszmary"[1]. Dotarliśmy

na Księżyc, okiełznaliśmy energię atomu, ale rasizm, chciwość i przemoc wciąż są powszechne. Nasz rozwój społeczny znajduje się daleko w tyle za naszym rozwojem technologicznym.

Jako filozofka muszę wyznać, że w mojej dziedzinie nie pojawiły się na razie żadne wiążące wnioski dotyczące natury umysłu, co może się wydać nieco mniej niepokojące niż powyższy problem. Jednak niezrozumienie pewnych zagadnień w filozofii również ma swoją cenę, jak się przekonamy podczas rozważań na dwa główne tematy niniejszej książki.

Pierwszy z nich jest nam dobrze znany. Obcujesz z nim przez całe swoje życie: to twoja świadomość. Zauważ, że gdy czytasz te słowa, masz poczucie samego lub samej siebie. Masz doznania zmysłowe, widzisz słowa na stronicy i tak dalej. Świadomość to odczuwana jakość twojego życia umysłowego. Gdyby nie ona, nie odczuwałbyś bólu, cierpienia, radości, ciekawości, żalu. Doświadczenia, pozytywne czy negatywne, zwyczajnie by nie istniały.

To świadomość każe ci pragnąć wakacji, wypraw do lasu czy niezwykłych posiłków. Ponieważ jest ona czymś tak bezpośrednim, tak dobrze znanym, w naturalny sposób rozumiesz ją za pośrednictwem twojego własnego przypadku. Nie musisz przecież czytać podręcznika do neurobiologii, by zrozumieć, w subiektywny sposób, jak to jest być świadomym. Świadomość to właśnie

to wewnętrzne odczucie. Ten rdzeń – nasze świadome doświadczenie – jest typową cechą posiadania świadomości.

Zła wiadomość: drugi główny temat książki wiąże się z rozpoznaniem, że filozoficzne implikacje sztucznej inteligencji mogą doprowadzić do pogorszenia się warunków życia świadomych bytów. Jeśli bowiem nie będziemy ostrożni, możemy doprowadzić do pojawienia się jednej lub wielu *przewrotnych konsekwencji* SI – do sytuacji, w której nie będzie ona ułatwiać nam życia, ale sprawi, że będziemy cierpieć, doprowadzi do naszej zagłady albo wyzysku innych świadomych bytów.

Wiele osób omawiało już zagrożenia dla rozwoju ludzkości, jakie może stworzyć SI. Obejmują one na przykład wyłączenie przez hakerów zasilania superinteligentnej broni autonomicznej wziętej prosto z filmu *Terminator*. Pytania, na których chcę się skupić, nie cieszyły się jednak aż tak wielką uwagą, choć są równie istotne. Przewrotne konsekwencje, które mam na myśli, można podzielić na dwie kategorie: 1) przeoczone sytuacje związane ze stworzeniem świadomych maszyn i 2) scenariusze dotyczące radykalnych ulepszeń mózgu, jak te wchodzące w skład oferty hipotetycznego Centrum Projektowania Umysłu. Rozpatrzmy teraz oba przypadki.

Świadome maszyny?

Załóżmy, że uda się nam stworzyć złożone sztuczne inteligencje o ogólnym zastosowaniu, takie, które mogą przechodzić od jednego zadania intelektualnego do drugiego i będą mogły dorównać ludziom pod względem zdolności umysłowych. Czy stworzymy w ten sposób *świadome* maszyny – maszyny będące jednostkami i podmiotami doświadczenia?

Jeśli chodzi o to, jak stworzyć maszynową świadomość, albo jakąkolwiek pewność czy w ogóle będziemy potrafili ją stworzyć, nie wiemy na razie nic. Jedno pozostaje jasne: to, czy SI będą zdolne do odczuwania subiektywnych doświadczeń, będzie kluczowe dla wartościowania przez nas ich istnienia. Świadomość to filozoficzny fundament naszych systemów etycznych, pozwalający nam określać, czy ktoś bądź coś jest podmiotem lub osobą, a nie tylko maszyną. Gdyby SI były świadome, zmuszanie ich do służenia nam byłoby równoznaczne z niewolnictwem. Czy mógłbyś spokojnie korzystać z usług sklepu z androidami, gdyby oferował on na sprzedaż świadome istoty, których zdolności umysłowe są równe ludzkim lub je przewyższają?

Gdybym była dyrektorką do spraw SI w Google czy Facebooku, myśląc o przyszłych projektach, wolałabym uniknąć etycznego zamieszania związanego

z nieumyślnym stworzeniem świadomego systemu. Mogłoby to doprowadzić do oskarżeń o zniewolenie SI i innych koszmarów wizerunkowych. Mogłoby nawet doprowadzić do zakazu stosowania technologii SI w niektórych sektorach.

Sądzę, że może to doprowadzić firmy produkujące SI do *inżynierii świadomości* – celowych działań mających na celu uniknięcie tworzenia świadomych SI do pewnych celów oraz tworzenie świadomych SI do innych celów, jeśli będzie to odpowiedni wybór. Oczywiście wiąże się to z założeniem, że świadomość to coś, co można zaprojektować albo wyeliminować. Być może świadomość będzie nieuniknionym skutkiem ubocznym stworzenia inteligentnego systemu; równie możliwe jest jednak to, że nigdy nie uda się jej stworzyć. W perspektywie długoterminowej sytuacja może się obrócić przeciwko ludziom: nie będzie już chodzić o to, że my możemy skrzywdzić SI, ale o to, że ona może skrzywdzić nas. Zdaniem niektórych inteligencja syntetyczna będzie nowym stadium ewolucji inteligencji na Ziemi. Ty i ja oraz nasz sposób doświadczania świata to tylko pośredni etap rozwoju SI, szczebel na ewolucyjnej drabinie. Na przykład Stephen Hawking, Nick Bostrom, Elon Musk, Max Tegmark, Bill Gates i wielu innych zwracają uwagę na „problem kontroli", który dotyczy sposobu kontrolowania SI przez ludzi, gdyby miała ona stać się od nich inteligentniejsza[2]. Przypuśćmy,

że stworzymy SI o inteligencji równej ludzkiej. Wyposażona w algorytmy samodoskonalenia się i moc szybkich obliczeń szybko znalazłaby sposób na zwiększenie poziomu swojej inteligencji, przekształcając się w superinteligencję – zaczęłaby przewyższać nas pod każdym względem. Zapewne wtedy stracilibyśmy nad nią kontrolę. Teoretycznie mogłaby ona doprowadzić do naszej zagłady. To tylko jeden ze sposobów, w jaki syntetyczne byty mogłyby zająć miejsce inteligencji organicznych; ludzie mogą też połączyć się z SI za pomocą coraz rozleglejszych udoskonaleń mózgu.

Problem kontroli zyskał nagłośnienie o skali światowej dzięki bestsellerowi Nicka Bostroma *Superinteligencja: scenariusze, strategie, zagrożenia* (Bostrom 2016). Nikt jednak nie zastanawia się nad tym, że świadomość może być kluczowa dla sposobu, w jaki SI będzie oceniać *nas*. Superinteligentna SI mogłaby, wykorzystując własne subiektywne doświadczenie jako punkt wyjścia, rozpoznać w nas zdolność świadomego doświadczania. W końcu my sami wartościujemy życie zwierząt innych niż my w zależności od tego, czy dostrzegamy w nich świadome byty – większość z nas wzdragałaby się przez zabiciem szympansa, ale nie przed zjedzeniem pomarańczy. Gdyby zaś inteligentne maszyny nie były świadome, czy to z powodu, że nie mogłyby posiadać świadomości, czy dlatego, że tak by je zaprojektowano, mogłoby to oznaczać dla nas kłopoty.

Ważne jest, by umieścić te zagadnienia w szerszym, obejmującym cały wszechświat kontekście. Gdy realizowałam dwuletni projekt w NASA, postawiłam tezę, że podobne procesy mogą zachodzić również na innych planetach; gdzieś w kosmosie inne gatunki mogły zostać zastąpione przez inteligencje syntetyczne. Poszukując życia poza Ziemią, musimy brać pod uwagę to, że największe kosmiczne inteligencje mogą być *postbiologiczne* – mogą być sztucznymi inteligencjami, które wyewoluowały z biologicznych cywilizacji. A gdyby te sztuczne inteligencje były pozbawione świadomości, zastąpienie przez nie inteligencji biologicznych oznaczałoby, że wszechświat mógł (lub może) zostać pozbawiony populacji świadomych istot.

Jeśli kwestia świadomości SI jest tak ważna, jak twierdzę, powinniśmy móc się dowiedzieć, czy można ją stworzyć i czy my, Ziemianie, ją stworzyliśmy. W następnych rozdziałach opiszę różne sposoby określania, czy syntetyczna świadomość istnieje, przedstawiając testy, które stworzyłam podczas pracy w Instytucie Studiów Zaawansowanych w Princeton.

Teraz chciałabym rozważyć sugestię, że ludzie powinni połączyć się z SI. Przypuśćmy, że znajdujesz się w Centrum Projektowania Umysłu. Jakie, jeśli w ogóle, ulepszenia umysłu zamówisz z menu? Z pewnością zaczynasz już zdawać sobie sprawę, że decyzje o projektowaniu umysłu nie są wcale proste.

Czy mógłbyś połączyć się z SI?

Nie byłabym zaskoczona, gdybyś był równie zaniepokojony ideą udoskonalania mózgu za pomocą mikrochipów jak ja. Gdy piszę te słowa, programy na moim smartfonie prawdopodobnie śledzą moją lokalizację, słuchają mojego głosu, zapisują historię moich wyszukiwań i sprzedają te informacje reklamodawcom. Wydaje mi się, że wyłączyłam te funkcje, ale firmy tworzące aplikacje nadały temu procesowi tak nieprzejrzystą postać, że nie mogę być tego pewna. Jeśli firmy wytwarzające SI nie potrafią dziś choćby szanować naszej prywatności, wyobraźmy sobie, jaki potencjał nadużyć pojawi się, kiedy nasze najskrytsze myśli zostaną zakodowane na mikrochipach, a może nawet staną się dostępne w internecie.

Załóżmy jednak, że w dziedzinie regulacji dotyczących SI nastąpi postęp, a nasze mózgi będą chronione przed hakerami i chciwością korporacji. Możliwe, że wtedy zaczniesz odczuwać potrzebę dokonania ulepszeń, zwłaszcza że wszyscy wokół ciebie będą odnosić korzyści z tej technologii[3]. W końcu, jeśli połączenie się z SI oznacza superinteligencję i długowieczność, czy nie jest ono czymś lepszym od alternatywy – nieuniknionego rozpadu mózgu i ciała?

Idea, że ludzie powinni połączyć się z SI, jest dziś coraz popularniejsza; przedstawia się ją zarówno jako sposób na to, by pracownicy nie zostali zastąpieni przez sztuczną inteligencję, jak i jako drogę wiodącą ku superinteligencji i nieśmiertelności. Na przykład Elon Musk stwierdził niedawno, że ludzie mogą uniknąć zastąpienia przez SI „dzięki swego rodzaju fuzji inteligencji biologicznej i inteligencji maszynowej" (Solon 2017). By zrealizować ten cel, założył nową firmę, Neuralink. Jednym z pierwszych jej dążeń jest stworzenie „koronki neuronalnej", wstrzykiwanej sieci łączącej mózg bezpośrednio z komputerami. Koronka neuronalna i inne oparte na SI udoskonalenia mają pozwolić na przekazywanie danych z mózgu do urządzeń cyfrowych albo chmury, gdzie dostępna jest potężna moc obliczeniowa.

Musk może się jednak kierować motywami innymi niż czysto altruistyczne. Promuje on linię produktów z dziedziny SI mających rozwiązać problem, który stworzyła sama ta dziedzina. Być może ulepszenia te okażą się korzystne, jednak by przekonać się, czy tak będzie, musimy wykroczyć poza cały ten medialny szum. Politycy, społeczeństwo, a nawet badacze SI potrzebują lepszego pojęcia o tym, jakie są stawki tej zmiany.

Jeśli na przykład SI nie może stać się świadoma, to zastępując części mózgu odpowiedzialne za świadomość mikrochipem, przestałbyś istnieć jako istota świadoma.

Stałbyś się czymś, co filozofowie nazywają zombie – pozbawionym świadomości symulakrum twojego wcześniejszego ja. Co więcej, nawet gdyby implanty te mogły zastąpić wytwarzające świadomość części mózgu bez przekształcania cię w zombie, radykalne ulepszanie wciąż wiąże się z poważnym ryzykiem – po wprowadzeniu zbyt wielu zmian może się okazać, że powstała w ten sposób osoba nie jest w ogóle tobą. Każdy człowiek decydujący się na ulepszanie może nieświadomie zakończyć podczas tego procesu własne życie.

Moim zdaniem wielu zwolenników radykalnego ulepszania nie uwzględnia tego, że ulepszona istota może nie być już tą, która się na ulepszenia zdecydowała. Na ogół przekonuje ich pogląd głoszący, że umysł jest programem komputerowym. Ich zdaniem można znacząco ulepszyć sprzęt, czyli mózg, i uruchomić na nim ten sam program, a zatem uznać, że umysł wciąż istnieje. Tak jak można wysłać i pobrać plik, tak umysł, pojmowany jako program, można załadować do chmury. Oto droga do nieśmiertelności w ujęciu technofila – nowe „życie po życiu" dla umysłu, który żyłby dłużej niż ciało. Choć technologiczna nieśmiertelność może jawić się jako kusząca, postaram się dowieść, że ten pogląd na umysł jest bardzo niedoskonały.

Jeśli więc za kilkadziesiąt lat trafisz do Centrum Projektowania Umysłu albo sklepu z androidami, pamiętaj, że to, co kupisz, może nie spełnić swojego

zadania ze złożonych przyczyn filozoficznych. *Niech kupujący się strzeże*. Zanim jednak zagłębimy się w tę problematykę, zgłosisz pewnie zarzut, że wszystkie te rozważania są hipotetyczne, ponieważ błędnie zakładam, że można stworzyć złożoną SI. Na jakiej podstawie uznaję, że te zmiany mogą nastąpić?

Rozdział 1

Epoka sztucznej inteligencji

Możliwe, że nie myślisz o SI na co dzień, ale jest ona wszędzie wokół ciebie. Jest obecna, kiedy wyszukujesz coś w Google. Była obecna, pokonując mistrzów świata w *Jeopardy!* i go. Co więcej, z każdą chwilą staje się ona doskonalsza. Nie istnieje jednak jeszcze sztuczna inteligencja o ogólnym zastosowaniu – taka, która potrafi prowadzić inteligentną rozmowę, integrować idee dotyczące różnych zagadnień, a nawet, być może, myśleć lepiej niż ludzie. SI tego rodzaju pojawia się w filmach takich jak *Ona* czy *Ex Machina*, może więc sprawiać wrażenie czegoś rodem z science fiction.

Podejrzewam jednak, że wcale nie jest ona tak odległa. Rozwojem SI rządzą siły rynku i przemysł obronny – miliardy dolarów przeznacza się dziś na konstruowanie inteligentnych asystentów domowych, robotycznych superżołnierzy i superkomputerów, których działanie