Etyka sztucznej inteligencji

Możliwość tworzenia myślących maszyn rodzi szereg kwestii etycznych, związanych zarówno z zapewnieniem, że maszyny te nie wyrządzą krzywdy ludziom i innym istotom istotnym moralnie, jak i ze statusem moralnym samych maszyn. W tym rozdziale omówiono niektóre wyzwania etyczne, które mogą się pojawić podczas tworzenia sztucznej inteligencji różnego rodzaju i stopnia.

Etyka w uczeniu maszynowym i innych algorytmach AI specyficznych dla danej dziedziny

Wyobraź sobie, że w niedalekiej przyszłości bank będzie korzystał z algorytmu uczenia maszynowego, aby rekomendować wnioski o kredyt hipoteczny do zatwierdzenia. Odrzucony wniosek wnosi pozew przeciwko bankowi, twierdząc, że algorytm dyskryminuje rasowo osoby ubiegające się o kredyt hipoteczny. Bank odpowiada, że jest to niemożliwe, gdyż algorytm celowo jest zaślepiony na rasę wnioskodawców. Istotnie, było to jednym z powodów, dla których bank zdecydował się na wdrożenie systemu. Mimo to statystyki pokazują, że wskaźnik akceptacji banku dla czarnoskórych kandydatów stale spada. Złożenie dziesięciu pozornie równie wykwalifikowanych, prawdziwych kandydatów (co ustali oddzielny panel składający się z ludzi) pokazuje, że algorytm akceptuje kandydatów rasy białej i odrzuca kandydatów rasy czarnej. Co może się dziać? Znalezienie odpowiedzi może nie być łatwe. Jeśli algorytm uczenia maszynowego opiera się na skomplikowanej sieci neuronowej lub algorytmie genetycznym powstałym w wyniku ukierunkowanej ewolucji, zrozumienie, dlaczego lub nawet w jaki sposób algorytm ocenia kandydatów na podstawie ich rasy, może okazać się prawie niemożliwe. Z drugiej strony uczenie maszynowe oparte na drzewach decyzyjnych lub sieciach bayesowskich jest znacznie bardziej przejrzyste dla inspekcji programisty, co może umożliwić snu audytor odkrył, że algorytm sztucznej inteligencji wykorzystuje dane adresowe wnioskodawców, którzy urodzili się lub wcześniej mieszkali na obszarach przeważnie dotkniętych ubóstwem. Algorytmy sztucznej inteligencji odgrywają coraz większą rolę we współczesnym społeczeństwie, chociaż zwykle nie są określane mianem „AI". Scenariusz opisany powyżej może mieć miejsce już w chwili, gdy piszemy. Coraz ważniejsze będzie rozwijanie algorytmów sztucznej inteligencji, które będą nie tylko wydajne i skalowalne, ale także przejrzyste dla kontroli – by wymienić tylko jedną z wielu ważnych społecznie właściwości. Niektóre wyzwania związane z etyką maszyn są podobne do wielu innych wyzwań związanych z projektowaniem maszyn. Zaprojektowanie ramienia robota, aby uniknąć zmiażdżenia bezpańskich ludzi, nie jest bardziej obciążone moralnie niż zaprojektowanie sofy trudnopalnej. Wiąże się to z nowymi wyzwaniami programistycznymi, ale nie wiąże się z nowymi wyzwaniami etycznymi. Kiedy jednak algorytmy sztucznej inteligencji podejmują się pracy poznawczej o wymiarze społecznym – zadań poznawczych wykonywanych wcześniej przez człowieka – algorytm sztucznej inteligencji przejmuje wymagania społeczne. Z pewnością frustrujące byłoby stwierdzenie, że żaden bank na świecie nie zatwierdzi Twojego pozornie doskonałego wniosku o pożyczkę i nikt nie wie dlaczego i nikt nie może się tego dowiedzieć, nawet w zasadzie. (Może masz imię silnie kojarzące się z deadbeatami? Kto wie?) Przejrzystość nie jest jedyną pożądaną cechą AI. Ważne jest także, aby algorytmy AI przejmujące funkcje społeczne były przewidywalne dla tych, którymi rządzą. Aby zrozumieć znaczenie takiej przewidywalności, rozważmy analogię. Zasada prawna stare decisis zobowiązuje sędziów do stosowania się do precedensów z przeszłości, gdy tylko jest to możliwe. Inżynierowi ta preferencja precedensu może wydawać się niezrozumiała – po co wiązać przyszłość z przeszłością, skoro technologia stale się rozwija? Ale jedną z najważniejszych funkcji systemu prawnego jest przewidywalność, tak aby można było np. spisywać umowy, wiedząc, w jaki sposób zostaną wykonane. Zadaniem systemu prawnego nie jest koniecznie optymalizacja społeczeństwa, ale zapewnienie przewidywalnego środowiska, w którym obywatele mogą optymalizować swoje życie. Coraz ważniejsze będzie również, aby algorytmy sztucznej inteligencji były odporne na manipulację. System wizyjny skanujący bagaż lotniczy w poszukiwaniu bomb musi być odporny na ludzkich przeciwników, którzy

celowo szukają możliwych do wykorzystania błędów w algorytmie – na przykład kształtu, który umieszczony obok pistoletu w bagażu uniemożliwiłby jego rozpoznanie. Odporność na manipulację jest zwyczajnym kryterium bezpieczeństwa informacji – niemal kryterium. Nie jest to jednak kryterium, które często pojawia się w czasopismach zajmujących się uczeniem maszynowym, które obecnie bardziej interesują się na przykład tym, jak algorytm skaluje się w większych systemach równoległych. Kolejnym ważnym kryterium społecznym w kontaktach z organizacjami jest możliwość znalezienia osoby odpowiedzialnej za wykonanie jakiegoś zadania. Kto ponosi winę, gdy system sztucznej inteligencji nie wykona powierzonego mu zadania? Programiści? Użytkownicy końcowi? Współcześni biurokraci często uciekają się do ustalonych procedur, które rozdzielają odpowiedzialność tak szeroko, że nie można wskazać jednej osoby winnej wynikających z tego katastrof . Jeszcze lepszym schronieniem może się okazać bezinteresowny osąd systemu eksperckiego. Nawet jeśli system sztucznej inteligencji został zaprojektowany tak, aby użytkownik mógł go obejść, należy wziąć pod uwagę motywację do kariery biurokraty, który będzie osobiście obwiniany, jeśli obejście się nie powiedzie, a który zdecydowanie wolałby obwiniać sztuczną inteligencję za każdą trudną decyzję o negatywnym wyniku. Odpowiedzialność, przejrzystość, kontrolowalność, nieprzekupność, przewidywalność i tendencja do nie wywoływania krzyku bezbronnej frustracji niewinnych ofiar: wszystkie kryteria odnoszące się do ludzi pełniących funkcje społeczne; wszystkie kryteria, które należy uwzględnić w algorytmie mającym zastąpić ludzką ocenę funkcji społecznych; wszystkie kryteria, które mogą nie pojawić się w czasopiśmie poświęconym uczeniu maszynowemu, biorąc pod uwagę sposób skalowania algorytmu do większej liczby komputerów. Ta lista kryteriów nie jest w żadnym wypadku wyczerpująca, ale służy jako mała próbka tego, o czym powinno myśleć coraz bardziej skomputeryzowane społeczeństwo.

Sztuczna inteligencja ogólna

Wśród współczesnych specjalistów zajmujących się sztuczną inteligencją panuje niemal powszechna zgoda co do tego, że sztuczna inteligencja w pewnym krytycznym sensie nie spełnia ludzkich możliwości, mimo że algorytmy sztucznej inteligencji pokonały ludzi w wielu konkretnych dziedzinach, takich jak szachy. Niektórzy sugerują, że gdy tylko badacze sztucznej inteligencji wymyślą, jak coś zrobić, zdolność ta przestaje być uważana za inteligentną – szachy były uważane za uosobienie inteligencji, dopóki Deep Blue nie zdobył mistrzostwa świata z rąk Kasparowa – ale nawet ci badacze są zgodni że współczesnym AI brakuje czegoś ważnego. Chociaż ta dziedzina sztucznej inteligencji dopiero się łączy, „sztuczna inteligencja ogólna" (zwana dalej AGI) to wyłaniający się termin artystyczny używany do określenia „prawdziwej" sztucznej inteligencji. Jak sama nazwa wskazuje, wyłaniający się konsensus jest taki, że brakującą cechą jest ogólność. Obecne algorytmy sztucznej inteligencji o wydajności porównywalnej z ludzką lub wyższą od ludzkiej charakteryzują się celowo zaprogramowanymi kompetencjami tylko w jednej, ograniczonej domenie. Deep Blue został mistrzem świata w szachach, ale nie potrafi nawet grać w warcaby, nie mówiąc już o prowadzeniu samochodu czy dokonaniu odkrycia naukowego. Takie nowoczesne algorytmy sztucznej inteligencji przypominają całe życie biologiczne z jedynym wyjątkiem Homo sapiens. Pszczoła wykazuje umiejętność budowania uli; bóbr wykazuje umiejętność budowania tam; ale pszczoła nie buduje tam, a bóbr nie nauczy się budować ula. Człowiek, obserwując, może nauczyć się jednego i drugiego; jest to jednak wyjątkowa zdolność wśród biologicznych form życia. Dyskusyjne jest, czy inteligencja ludzka jest rzeczywiście powszechna – z pewnością w niektórych zadaniach poznawczych jesteśmy lepsi od innych – ale inteligencja ludzka z pewnością ma znacznie szersze zastosowanie niż inteligencja niehominidów. Stosunkowo łatwo jest przewidzieć, jakie problemy związane z bezpieczeństwem mogą wynikać z działania sztucznej inteligencji tylko w określonej domenie. Radzenie sobie z AGI działającym w wielu nowych kontekstach, których nie można przewidzieć z góry, stanowi jakościowo inną klasę problemów. Kiedy inżynierowie budują reaktor jądrowy, wyobrażają sobie konkretne zdarzenia, które mogą mieć

w nim miejsce – awarie zaworów, awarie komputerów, wzrost temperatury rdzeni – i projektują reaktor tak, aby zdarzenia te nie były katastrofalne. Lub, na bardziej przyziemnym poziomie, zbudowanie tostera polega na wyobrażeniu sobie chleba i wyobrażeniu sobie reakcji chleba na element grzejny tostera. Sam toster nie wie, że jego celem jest zrobienie tostów – cel tostera jest przedstawiony w umyśle projektanta, ale nie jest wyraźnie przedstawiony w obliczeniach wewnątrz tostera – więc jeśli włożysz szmatkę do tostera, może się ona zaczepić ogień, ponieważ projekt jest realizowany w nieprzewidzianym kontekście z nieprzewidzianymi efektami ubocznymi. Nawet algorytmy sztucznej inteligencji specyficzne dla danego zadania wyrzucają nas poza paradygmat tostera, czyli domenę lokalnie zaprogramowanego, specjalnie zaplanowanego zachowania. Weźmy pod uwagę Deep Blue, algorytm szachowy, który pokonał Garriego Kasparowa w walce o mistrzostwo świata w szachach. Gdyby było tak, że maszyny mogą robić tylko dokładnie to, co im się każe, programiści musieliby ręcznie zaprogramować bazę danych zawierającą ruchy dla każdej możliwej pozycji szachowej, jaką Deep Blue mógłby napotkać. Ale dla programistów Deep Blue nie była to opcja. Po pierwsze, przestrzeń możliwych pozycji szachowych jest niewyobrażalnie duża. Po drugie, gdyby programiści w każdej możliwej sytuacji ręcznie wprowadzili to, co uznali za dobre posunięcie, powstały system nie byłby w stanie wykonywać silniejszych ruchów szachowych niż jego twórcy. Ponieważ sami programiści nie byli mistrzami świata, taki system nie byłby w stanie pokonać Garry'ego Kasparowa.

Tworząc nadludzkiego szachistę, ludzcy programiści z konieczności poświęcili swoją zdolność przewidywania lokalnego, specyficznego zachowania Deep Blue w grze. Zamiast tego programiści Deep Blue mieli (uzasadnioną) pewność, że ruchy szachowe Deep Blue spełnią nielokalne kryterium optymalności: mianowicie, że ruchy te będą miały tendencję do kierowania przyszłością planszy na wyniki w „zwycięskim" regionie zgodnie z definicją według zasad szachowych. Ta prognoza dotycząca odległych konsekwencji, choć okazała się trafna, nie pozwoliła programistom wyobrazić sobie lokalnego zachowania Deep Blue – jego reakcji na konkretny atak na swojego króla – ponieważ Deep Blue obliczył nielokalną mapę gry, połączenie pomiędzy ruchu i jego możliwych przyszłych konsekwencji dokładniej, niż mogliby to zrobić programiści. Współcześni ludzie robią dosłownie miliony rzeczy, aby się wyżywić – aby służyć ostatecznym konsekwencjom bycia nakarmionym. Niewiele z tych działań zostało „przewidzianych przez Naturę" w sensie wyzwań przodków, do których jesteśmy bezpośrednio przystosowani. Jednak nasz zaadaptowany mózg stał się na tyle potężny, że można go było zastosować w znacznie szerszym zakresie; abyśmy mogli przewidzieć konsekwencje milionów różnych działań w różnych dziedzinach i określić nasze preferencje co do ostatecznych wyników. Ludzie przemierzyli przestrzeń kosmiczną i pozostawili ślady na Księżycu, mimo że żaden z naszych przodków nie spotkał się z wyzwaniem analogicznym do próżni. W porównaniu ze sztuczną inteligencją specyficzną dla domeny zaprojektowanie systemu, który będzie bezpiecznie działał w tysiącach kontekstów, stanowi jakościowo inny problem; włączając konteksty, które nie zostały specjalnie przewidziane ani przez projektantów, ani przez użytkowników; włączając konteksty, z którymi nie spotkał się jeszcze żaden człowiek. Być może nie ma tutaj lokalnej specyfikacji dobrego zachowania – nie ma prostej specyfikacji samych zachowań, tak samo jak nie istnieje zwięzły lokalny opis wszystkich sposobów, w jakie ludzie zdobywają chleb powszedni. Aby zbudować sztuczną inteligencję, która będzie działać bezpiecznie w wielu obszarach, z wieloma konsekwencjami, w tym z problemami, których inżynierowie nigdy wyraźnie nie przewidywali, należy określić dobre zachowanie w taki sposób, jak „X takie, że konsekwencja X nie będzie szkodliwa dla ludzi". To nie jest lokalne; polega na ekstrapolowaniu odległych konsekwencji działań. Zatem jest to tylko skuteczna specyfikacja – taka, która może zostać zrealizowana jako właściwość projektowa – jeśli system wyraźnie ekstrapoluje konsekwencje swojego zachowania. Toster nie może mieć tej właściwości konstrukcyjnej, ponieważ toster nie jest w stanie przewidzieć konsekwencji opiekania chleba. Wyobraźcie sobie inżyniera, który musi powiedzieć: „No cóż, nie mam pojęcia, jak ten samolot, który zbudowałem, będzie

bezpiecznie latał – w istocie nie mam pojęcia, jak w ogóle będzie latał, czy będzie trzepotał skrzydłami, czy napompuje się helem, czy czymś innym. nawet sobie tego nie wyobrażałem – ale zapewniam, że konstrukcja jest bardzo, bardzo bezpieczna." Może się to wydawać stanowiskiem nie do pozazdroszczenia z punktu widzenia public relations, ale trudno wyobrazić sobie, jaka inna gwarancja etycznego postępowania byłaby możliwa dla inteligencji ogólnej, operującej na nieprzewidzianych problemach, w różnych dziedzinach, z preferencją nad odległymi konsekwencjami. Sprawdzenie projektu poznawczego może zweryfikować, czy umysł rzeczywiście poszukiwał rozwiązań, które sklasyfikowalibyśmy jako etyczne; ale nie mogliśmy przewidzieć, jakie konkretne rozwiązanie odkryje umysł. Przestrzeganie takiej weryfikacji wymaga pewnego sposobu odróżnienia wiarygodnych zapewnień (procedura, która nie powie, że sztuczna inteligencja jest bezpieczna, jeśli sztuczna inteligencja naprawdę nie jest bezpieczna) od czystej nadziei i magicznego myślenia („Nie mam pojęcia, w jaki sposób Kamień Filozoficzny przemieni ołów w złoto , ale zapewniam, że tak będzie!"). Należy pamiętać, że oczekiwania czysto pełne nadziei były wcześniej problemem w badaniach nad sztuczną inteligencją (McDermott 1976). Zbudowanie godnego zaufania AGI będzie wymagało innych metod i innego sposobu myślenia, począwszy od sprawdzania oprogramowania elektrowni pod kątem błędów – będzie wymagało AGI, które myśli jak ludzki inżynier dbający o etykę, a nie tylko jako prosty produkt inżynierii etycznej. Zatem dyscyplina etyki AI, szczególnie w zastosowaniu do AGI, prawdopodobnie będzie zasadniczo różnić się od dyscypliny etycznej technologii niekognitywnych w tym sensie, że:

* Lokalne, specyficzne zachowanie sztucznej inteligencji może nie być przewidywalne poza jej bezpieczeństwem, nawet jeśli programiści zrobią wszystko dobrze.

* Weryfikacja bezpieczeństwa systemu staje się większym wyzwaniem, ponieważ musimy zweryfikować, co system próbuje zrobić, zamiast być w stanie zweryfikować bezpieczne zachowanie systemu we wszystkich kontekstach operacyjnych.

* Samo poznanie etyczne należy traktować jako przedmiot inżynierii

Maszyny ze statusem moralnym

Inny zestaw kwestii etycznych pojawia się, gdy rozważamy możliwość, że niektóre przyszłe systemy sztucznej inteligencji mogą kandydować do posiadania statusu moralnego. Nasze postępowanie z istotami posiadającymi status moralny nie jest wyłącznie kwestią racjonalności instrumentalnej. Mamy także moralne powody, aby traktować je w określony sposób i powstrzymywać się od traktowania ich w określony inny sposób. Francis Kamm zaproponował następującą definicję statusu moralnego, która będzie nam służyć:

X ma status moralny = ponieważ X liczy się moralnie jako taki, dopuszczalne/niedopuszczalne jest robienie z nim rzeczy dla niego samego.

(Parafraza z Kamm 2007)

Skała nie ma statusu moralnego. Możemy ją zmiażdżyć, sproszkować lub poddać dowolnej obróbce, bez obawy o samą skałę. Człowieka zaś należy traktować nie tylko jako środek, ale i cel. Różne teorie etyczne nie zgadzają się dokładnie z tym, co oznacza traktowanie osoby jako celu; ale z pewnością wiąże się to z wzięciem pod uwagę jej uzasadnionych interesów – zwróceniem uwagi na jej dobro – i może również wiązać się z akceptacją surowych moralnych ograniczeń ubocznych w naszych kontaktach z nią, takich jak zakaz mordowania jej, okradania jej lub wykonywania różnych innych rzeczy na jej własność bez jej zgody. Co więcej, dlatego, że osoba ludzka liczy się sama w sobie i ze względu na nią, niedopuszczalne jest czynić jej takie rzeczy. Można to wyrazić bardziej zwięźle, stwierdzając, że osoba ludzka ma status moralny. Pytania o status moralny są ważne w niektórych

obszarach etyki praktycznej. Na przykład spory dotyczące moralnej dopuszczalności aborcji często opierają się na nieporozumieniach dotyczących moralnego statusu embrionu. Kontrowersje wokół eksperymentów na zwierzętach i traktowania zwierząt w przemyśle spożywczym wiążą się z pytaniami o status moralny różnych gatunków zwierząt. Nasze zobowiązania wobec ludzi z ciężką demencją, takich jak pacjenci z chorobą Alzheimera w późnym stadium, mogą również zależeć od kwestii statusu moralnego. Powszechnie uważa się, że obecne systemy sztucznej inteligencji nie mają statusu moralnego. Możemy zmieniać, kopiować, kończyć, usuwać lub używać programów komputerowych według własnego uznania; przynajmniej jeśli chodzi o same programy. Wszystkie ograniczenia moralne, jakim podlegamy w kontaktach ze współczesnymi systemami sztucznej inteligencji, opierają się na naszych obowiązkach wobec innych istot, takich jak nasi bliźni, a nie na jakichkolwiek obowiązkach wobec samych systemów. Chociaż powszechnie uważa się, że współczesnym systemom sztucznej inteligencji brakuje statusu moralnego, nie jest jasne, jakie dokładnie atrybuty wpływają na ten status moralny. Powszechnie proponuje się dwa kryteria jako istotnie powiązane ze statusem moralnym, osobno lub w połączeniu: zdolność odczuwania i rozumność (lub osobowość). Można je scharakteryzować z grubsza następująco:

Wrażliwość: zdolność do fenomenalnych doświadczeń lub jakości, takich jak zdolność odczuwania bólu i cierpienia.

Rozumność: zestaw zdolności związanych z wyższą inteligencją, takich jak samoświadomość i zdolność reagowania na rozum.

Jednym z powszechnych poglądów jest to, że wiele zwierząt posiada qualia i dlatego ma pewien status moralny, ale tylko istoty ludzkie posiadają mądrość, co daje im wyższy status moralny niż zwierzęta inne niż ludzie.1 Pogląd ten musi oczywiście stawić czoła istnieniu granic przypadki takie jak z jednej strony niemowlęta ludzkie lub istoty ludzkie ze znacznym upośledzeniem umysłowym – czasami niestety nazywane „ludźmi marginalnymi" – które nie spełniają kryteriów mądrości; z drugiej strony niektóre zwierzęta inne niż ludzie, takie jak wielkie małpy człekokształtne, które mogą posiadać przynajmniej niektóre elementy mądrości. Niektórzy zaprzeczają, że tak zwani „ludzie marginalni" mają pełny status moralny. Inni proponują dodatkowe sposoby, w jakie przedmiot mógłby kwalifikować się jako posiadacz statusu moralnego, na przykład poprzez bycie członkiem rodzaju, który normalnie posiada zdolność odczuwania lub rozumu, lub poprzez pozostawanie w odpowiedniej relacji z jakąś istotą, która niezależnie ma status moralny . Jednakże dla obecnych celów skupimy się na kryteriach odczuwania i rozsądku. Ten obraz statusu moralnego sugeruje, że system sztucznej inteligencji będzie miał pewien status moralny, jeśli będzie miał zdolność do tworzenia qualiów, takich jak zdolność odczuwania bólu. Czujący system sztucznej inteligencji, nawet jeśli brakuje mu języka i innych wyższych zdolności poznawczych, nie przypomina pluszowego zwierzaka czy nakręcanej lalki; bardziej przypomina żywe zwierzę. Zadawanie bólu myszy jest niewłaściwe, chyba że istnieją ku temu wystarczająco mocne, nadrzędne powody moralne. To samo dotyczy każdego świadomego systemu sztucznej inteligencji. Jeżeli oprócz zdolności odczuwania system sztucznej inteligencji posiada także rozum podobny do tej, jaki posiada normalny dorosły człowiek, wówczas miałby pełny status moralny, równy statusowi istoty ludzkiej. Jedną z idei leżących u podstaw tej oceny moralnej można wyrazić w mocniejszej formie jako zasada niedyskryminacji:

Zasada niedyskryminacji podłoża: Jeżeli dwie istoty mają tę samą funkcjonalność i to samo świadome doświadczenie, a różnią się jedynie podłożem ich realizacji, to mają ten sam status moralny.

Można argumentować za tą zasadą, argumentując, że odrzucenie jej byłoby równoznaczne z przyjęciem stanowiska podobnego do rasizmu. Podłoże nie ma fundamentalnego znaczenia moralnego w taki sam sposób i z tego samego powodu, co kolor skóry. Zasada niedyskryminacji podłoża nie

oznacza, że komputer cyfrowy może być świadomy ani że może mieć taką samą funkcjonalność jak istota ludzka. Podłoże może oczywiście mieć znaczenie moralne, o ile ma wpływ na zdolność odczuwania lub funkcjonalność. Ale utrzymując te rzeczy na stałym poziomie, nie ma moralnej różnicy, czy istota jest zbudowana z krzemu, czy z węgla, i czy jej mózg wykorzystuje półprzewodniki lub neuroprzekaźniki. Można również zaproponować, że fakt, że systemy sztucznej inteligencji są sztuczne – tj. stanowią produkt celowego projektu – nie ma zasadniczo znaczenia dla ich statusu moralnego. Zasadę tę moglibyśmy sformułować następująco:

Zasada niedyskryminacji ontogenezy: Jeśli dwie istoty mają tę samą funkcjonalność i to samo doświadczenie świadomości, a różnią się jedynie sposobem powstania, wówczas mają ten sam status moralny.

Dziś koncepcja ta jest powszechnie akceptowana w przypadku człowieka, chociaż w niektórych kręgach, szczególnie w przeszłości, wpływowy był pogląd, że status moralny danej osoby zależy od jej rodu lub kasty. Nie wierzymy, że czynniki przyczynowe, takie jak planowanie rodziny, poród wspomagany, zapłodnienie in vitro, selekcja gamet, celowa poprawa odżywiania matki itd. – które wprowadzają element świadomego wyboru i projektu w tworzeniu osób ludzkich – mają jakikolwiek wpływ niezbędne implikacje dla statusu moralnego potomstwa. Nawet ci, którzy sprzeciwiają się klonowaniu reprodukcyjnemu ludzi ze względów moralnych lub religijnych, generalnie zgadzają się, że w przypadku doprowadzenia do porodu ludzkiego klonu miałby on taki sam status moralny jak każde inne ludzkie dziecko. Zasada niedyskryminacji ontogenezy rozszerza to rozumowanie na przypadek dotyczący całkowicie sztucznych systemów poznawczych. Jest oczywiście możliwe, że okoliczności stworzenia wpłyną na powstałe potomstwo w taki sposób, że zmienią jego status moralny. Na przykład, jeśli podczas poczęcia lub ciąży przeprowadzono jakąś procedurę, która spowodowała rozwój płodu ludzkiego bez mózgu, wówczas fakt dotyczący ontogenezy miałby znaczenie dla naszej oceny statusu moralnego potomstwa. Jednakże dziecko bezmózgowe miałoby taki sam status moralny jak każde inne podobne dziecko bezmózgowe, łącznie z dzieckiem, które powstało w wyniku całkowicie naturalnego procesu. Różnica w statusie moralnym między dzieckiem bezmózgim a dzieckiem normalnym opiera się na jakościowej różnicy między nimi – na fakcie, że jedno ma umysł, a drugie nie. Ponieważ dwoje dzieci nie ma tej samej funkcjonalności i tego samego świadomego doświadczenia, zasada niedyskryminacji ontogenezy nie ma zastosowania. Chociaż zasada niedyskryminacji ontogenezy stwierdza, że ontogeneza istoty nie ma istotnego wpływu na jej status moralny, nie zaprzecza, że fakty dotyczące ontogenezy mogą wpływać na obowiązki, jakie mają poszczególne podmioty moralne wobec danej istoty. Rodzice mają wobec swojego dziecka szczególne obowiązki, których nie mają wobec innych dzieci i których nie mieliby, nawet gdyby istniało drugie dziecko jakościowo identyczne z ich własnym. Podobnie zasada niedyskryminacji ontogenezy jest zgodna z twierdzeniem, że twórcy lub właściciele systemu sztucznej inteligencji posiadający status moralny mogą mieć szczególne obowiązki wobec swojego sztucznego umysłu, których nie mają wobec innego sztucznego umysłu, nawet jeśli dane umysły są jakościowo podobne i mają ten sam status moralny.

Jeśli przyjmiemy zasady niedyskryminacji ze względu na substrat i ontogenezę, wówczas na wiele pytań dotyczących tego, jak powinniśmy traktować sztuczne umysły, można odpowiedzieć, stosując te same zasady moralne, których używamy do określenia naszych obowiązków w bardziej znanych kontekstach. O ile obowiązki moralne wynikają ze względów dotyczących statusu moralnego, powinniśmy traktować sztuczny umysł w taki sam sposób, w jaki powinniśmy traktować jakościowo identyczny naturalny umysł ludzki w podobnej sytuacji. Upraszcza to problem opracowania etyki leczenia sztucznych umysłów. Nawet jeśli przyjmiemy to stanowisko, musimy jednak stawić czoła szeregowi nowych pytań etycznych, na które powyższe zasady pozostawiają bez odpowiedzi. Pojawiają się nowe pytania etyczne, ponieważ sztuczne umysły mogą mieć zupełnie inne właściwości niż zwykłe umysły ludzkie lub

zwierzęce. Musimy rozważyć, jak te nowe właściwości wpłyną na status moralny sztucznych umysłów i co oznaczałoby szanowanie statusu moralnego takich egzotycznych umysłów.

Umysły o egzotycznych właściwościach

W przypadku istot ludzkich zwykle nie wahamy się przypisywać wrażliwości i świadomego doświadczenia każdej jednostce, która wykazuje normalne ludzkie zachowanie. Niewielu wierzy, że istnieją inni ludzie, którzy zachowują się zupełnie normalnie, ale brakuje im świadomości. Jednakże inne istoty ludzkie nie tylko zachowują się w sposób podobny do nas samych; mają także mózgi i architekturę poznawczą zbudowaną podobnie jak my. Natomiast sztuczny intelekt może mieć zupełnie inną budowę niż ludzki intelekt, a mimo to nadal wykazywać ludzkie zachowanie lub posiadać skłonności behawioralne zwykle wskazujące na osobowość. Można zatem wyobrazić sobie sztuczny intelekt, który byłby rozumny i być może byłby osobą, ale nie byłby czujący ani nie miałby żadnego świadomego doświadczenia. (To, czy jest to naprawdę możliwe, zależy od odpowiedzi na pewne nietrywialne pytania metafizyczne.) Gdyby taki system był możliwy, pojawiłoby się pytanie, czy nie czująca osoba miałaby jakikolwiek status moralny; a jeśli tak, to czy miałby taki sam status moralny jak czująca osoba. Ponieważ zwykle przyjmuje się, że zdolność odczuwania lub przynajmniej zdolność odczuwania występuje u każdej jednostki będącej osobą, kwestii tej nie poświęcono dotychczas zbyt wiele uwagi2. Kolejna egzotyczna właściwość, która z pewnością jest metafizycznie i fizycznie możliwa dla jednostki sztuczna inteligencja, polega na tym, że jej subiektywne tempo czasu drastycznie odbiega od tempa charakterystycznego dla biologicznego mózgu człowieka. Koncepcję subiektywnego tempa czasu najlepiej wyjaśnić, wprowadzając najpierw ideę emulacji całego mózgu, czyli „przesyłania". „Przesyłanie" odnosi się do hipotetycznej przyszłej technologii, która umożliwi przeniesienie intelektu człowieka lub innego zwierzęcia z jego pierwotnej implementacji w organicznym mózgu na komputer cyfrowy. Jeden ze scenariuszy wygląda następująco: najpierw wykonywany jest skan konkretnego mózgu o bardzo wysokiej rozdzielczości, co prawdopodobnie powoduje zniszczenie oryginału. Na przykład mózg można zeszklić i podzielić na cienkie plasterki, które można następnie zeskanować za pomocą jakiejś formy wysokoprzepustowej mikroskopii połączonej z automatycznym rozpoznawaniem obrazu. Możemy sobie wyobrazić, że ten skan będzie wystarczająco szczegółowy, aby uchwycić wszystkie neurony, ich połączenia synaptyczne i inne cechy, które są funkcjonalnie istotne dla pierwotnego działania mózgu. Po drugie, tę trójwymiarową mapę składników mózgu i ich wzajemnych połączeń połączono z biblioteką zaawansowanej teorii neuronaukowej, która określa właściwości obliczeniowe każdego podstawowego typu elementu, takiego jak różne rodzaje neuronów i połączeń synaptycznych. Po trzecie, struktura obliczeniowa i związane z nią zachowanie algorytmiczne jej komponentów są zaimplementowane w jakimś potężnym komputerze. Jeśli proces przesyłania przebiegł pomyślnie, program komputerowy powinien teraz odtworzyć podstawowe cechy funkcjonalne pierwotnego mózgu. Powstały w ten sposób przesłany plik może znajdować się w symulowanej rzeczywistości wirtualnej lub, alternatywnie, może mieć kontrolę nad ciałem robota, umożliwiając mu bezpośrednią interakcję z zewnętrzną rzeczywistością fizyczną. W kontekście takiego scenariusza pojawia się wiele pytań: Na ile prawdopodobne jest, że pewnego dnia procedura ta stanie się wykonalna technologicznie? Gdyby procedura zadziałała i stworzyła program komputerowy wykazujący mniej więcej tę samą osobowość, te same wspomnienia i te same wzorce myślenia, co pierwotny mózg, czy ten program byłby świadomy? Czy przesyłana osoba będzie tą samą osobą, której mózg został rozebrany w procesie przesyłania? Co stanie się z tożsamością osobistą, jeśli przesyłany plik zostanie skopiowany w taki sposób, że równolegle będą działać dwa podobne lub jakościowo identyczne umysły przesyłające? Chociaż wszystkie te pytania są istotne dla etyki inteligencji maszynowej, skupmy się tutaj na kwestii związanej z pojęciem subiektywnego tempa czasu.

Załóżmy, że przesyłanie może być świadome. Jeśli uruchomimy program do przesyłania na szybszym komputerze, spowoduje to, że przesyłanie, jeśli będzie podłączone do urządzenia wejściowego, takiego jak kamera wideo, będzie odbierało świat zewnętrzny tak, jakby był spowolniony. Na przykład, jeśli przesyłanie przebiega tysiąc razy szybciej niż pierwotny mózg, wówczas świat zewnętrzny będzie wyglądał tak, jakby był tysiąckrotnie spowolniony. Ktoś upuszcza fizyczny kubek z kawą. Przesyłający obserwuje, jak kubek powoli spada na ziemię, podczas gdy przesyłany kończy czytać poranną gazetę i wysyła kilka e-maili. Jedna sekunda czasu obiektywnego odpowiada siedemnastu minutom czasu subiektywnego. Obiektywny i subiektywny czas trwania mogą zatem się różnić. Subiektywny czas to nie to samo, co ocena podmiotu lub postrzeganie szybkości jego upływu. Ludzie często mylą się co do upływu czasu. Możemy wierzyć, że jest godzina pierwsza, podczas gdy w rzeczywistości jest kwadrans po drugiej; lub lek pobudzający może sprawić, że nasze myśli będą galopować, sprawiając wrażenie, jakby upłynął czas w bardziej subiektywny sposób, niż jest to w rzeczywistości. Te przyziemne przypadki dotyczą raczej zniekształconego postrzegania czasu niż zmiany tempa subiektywnego czasu. Nawet w mózgu naćpanym kokainą prawdopodobnie nie następuje znacząca zmiana w szybkości wykonywania podstawowych obliczeń neurologicznych; bardziej prawdopodobne, że narkotyk powoduje szybsze przechodzenie mózgu z jednej myśli na drugą, przez co spędza on mniej subiektywnie czasu na myśleniu o większej liczbie odrębnych myśli. Zmienność subiektywnego tempa czasu jest egzotyczną właściwością sztucznych umysłów, która rodzi nowe problemy etyczne. Na przykład, w przypadkach, gdy czas trwania doświadczenia jest etycznie istotny, czy czas trwania należy mierzyć w czasie obiektywnym czy subiektywnym? Jeśli przesłany plik popełnił przestępstwo i został skazany na cztery lata więzienia, czy powinny to być cztery obiektywne lata – które mogą odpowiadać wielu tysiącleciom czasu subiektywnego – czy też powinny to być cztery subiektywne lata, które mogą minąć w ciągu kilku dni obiektywnego czasu? Jeśli szybka sztuczna inteligencja i człowiek odczuwają ból, czy pilniejsze jest złagodzenie bólu sztucznej inteligencji na tej podstawie, że odczuwa ona dłuższy subiektywny czas trwania bólu z każdą sekundą gwiezdną opóźnienia w uśmierzaniu bólu? Ponieważ w naszym zwyczajowym kontekście biologicznych ludzi czas subiektywny nie jest znacząco zmienny, nie jest zaskakujące, że tego rodzaju kwestii nie można bezpośrednio rozstrzygnąć za pomocą znanych norm etycznych, nawet jeśli normy te rozciągają się na sztuczne intelekty za pomocą zasad niedyskryminacji, takie jak te zaproponowane w poprzedniej sekcji. Aby zilustrować rodzaj twierdzeń etycznych, które mogą być tu istotne, formułujemy (ale nie argumentujemy) zasadę uprzywilejowującą czas subiektywny jako normatywnie bardziej fundamentalne pojęcie:

Zasada subiektywnej stopy czasu: W przypadkach, gdy czas trwania doświadczenia ma podstawowe znaczenie normatywne, liczy się subiektywny czas trwania doświadczenia.

Jak dotąd omówiliśmy dwie możliwości (nieświadoma świadomość i zmienna subiektywna szybkość czasu), które są egzotyczne w stosunkowo głębokim sensie metafizycznego problemu, a także pozbawione wyraźnych przykładów lub podobieństw we współczesnym świecie. Inne właściwości ewentualnych sztucznych umysłów byłyby egzotyczne w bardziej powierzchownym sensie; na przykład odbiegając w jakimś bezproblemowo ilościowym wymiarze od znanych nam rodzajów umysłu. Jednak takie powierzchownie egzotyczne właściwości mogą również stwarzać nowe problemy etyczne – jeśli nie na poziomie podstawowej filozofii moralności, to na poziomie etyki stosowanej lub zasad etycznych średniego szczebla. Jeden z ważnych zestawów egzotycznych właściwości sztucznej inteligencji wiąże się z reprodukcją. Szereg warunków empirycznych mających zastosowanie do reprodukcji ludzi nie musi mieć zastosowania do sztucznych inteligencji. Na przykład ludzkie dzieci są produktem rekombinacji materiału genetycznego obojga rodziców; rodzice mają ograniczoną możliwość wpływania na charakter swojego potomstwa; embrion ludzki musi rozwijać się w macicy przez dziewięć miesięcy; osiągnięcie dojrzałości przez ludzkie dziecko zajmuje piętnaście do dwudziestu lat; dziecko ludzkie nie dziedziczy umiejętności i wiedzy nabytych przez rodziców; istoty ludzkie posiadają złożony,

rozwinięty zestaw adaptacji emocjonalnych związanych z reprodukcją, wychowywaniem i relacją dziecko – rodzic. Żaden z tych warunków empirycznych nie musi odnosić się do kontekstu odtwarzającej się inteligencji maszyny. Jest zatem prawdopodobne, że wiele zasad moralnych średniego szczebla, które zaakceptowaliśmy jako normy regulujące reprodukcję człowieka, będzie wymagało ponownego przemyślenia w kontekście reprodukcji sztucznej inteligencji. Aby zilustrować, dlaczego niektóre z naszych norm moralnych wymagają ponownego przemyślenia w kontekście reprodukcji sztucznej inteligencji, wystarczy rozważyć tylko jedną egzotyczną właściwość sztucznej inteligencji: jej zdolność do szybkiej reprodukcji. Mając dostęp do sprzętu komputerowego, sztuczna inteligencja mogłaby bardzo szybko powielić się, w czasie nie dłuższym niż wykonanie kopii oprogramowania sztucznej inteligencji. Co więcej, ponieważ kopia AI byłaby identyczna z oryginałem, urodziłaby się w pełni dojrzała, a kopia mogłaby natychmiast rozpocząć tworzenie własnych kopii. W przypadku braku ograniczeń sprzętowych populacja sztucznej inteligencji mogłaby zatem rosnąć wykładniczo w niezwykle szybkim tempie, z czasem podwajania rzędu minut lub godzin, a nie dziesięcioleci lub stuleci. Nasze obecne normy etyczne dotyczące reprodukcji obejmują pewną wersję zasady wolności reprodukcyjnej, zgodnie z którą każda osoba lub para sama decyduje, czy i ile dzieci mieć. Inną normą, którą mamy (przynajmniej w krajach bogatych i o średnich dochodach) jest to, że społeczeństwo musi wkroczyć, aby zaspokoić podstawowe potrzeby dzieci w przypadkach, gdy ich rodzice nie są w stanie tego zrobić lub nie chcą tego zrobić. Łatwo dostrzec, jak te dwie normy mogłyby zderzyć się w kontekście bytów posiadających zdolność do niezwykle szybkiej reprodukcji. Weźmy pod uwagę na przykład populację przesłanych osób, z których jedna pragnie stworzyć jak największy klan. Biorąc pod uwagę całkowitą swobodę reprodukcji, przesłany plik może zacząć się kopiować tak szybko, jak to możliwe; a tworzone przez niego kopie – które mogą działać na nowym sprzęcie komputerowym będącym własnością oryginału lub przez niego wynajmowanym, lub mogą dzielić ten sam komputer co oryginał – również zaczną się kopiować, ponieważ są identyczne z przesłanym przodkiem i podzielają jego filoprogeniczne pragnienia. Wkrótce członkowie klanu przesyłającego nie będą w stanie opłacić rachunku za prąd ani czynszu za przetwarzanie obliczeniowe i przechowywanie potrzebne do utrzymania ich przy życiu. W tym momencie może włączyć się system opieki społecznej, który zapewni im przynajmniej to, co niezbędne do podtrzymania życia. Ale jeśli populacja będzie rosła szybciej niż gospodarka, zasoby się skończą; w tym momencie przesyłane pliki albo znikną, albo ich zdolność do reprodukcji zostanie ograniczona. Scenariusz ten ilustruje, jak niektóre zasady etyczne średniego szczebla, odpowiednie we współczesnych społeczeństwach, mogłyby wymagać modyfikacji, gdyby społeczeństwa te miały obejmować osoby posiadające egzotyczną cechę zdolności do bardzo szybkiej reprodukcji. Ogólna uwaga jest taka, że myśląc o etyce stosowanej w kontekstach bardzo różniących się od naszej znanej kondycji ludzkiej, musimy uważać, aby nie pomylić zasad etycznych średniego szczebla z podstawowymi prawdami normatywnymi. Inaczej mówiąc, musimy rozpoznać zarówno zakres, w jakim nasze zwykłe zasady normatywne są w sposób dorozumiany uwarunkowane uzyskaniem różnych warunków empirycznych, jak i potrzebę odpowiedniego dostosowania tych zasad przy stosowaniu ich do hipotetycznych futurystycznych przypadków, w których zakłada się, że ich warunki wstępne nie zostaną spełnione . Nie formułujemy w ten sposób żadnego kontrowersyjnego twierdzenia na temat relatywizmu moralnego, a jedynie podkreślamy zdroworozsądkowy pogląd, że kontekst jest istotny dla stosowania etyki – i sugerujemy, że ten punkt jest szczególnie istotny, gdy rozważa się etykę umysłów o egzotycznych właściwościach.

Superinteligencja

I. J. Good (1965) przedstawił klasyczną hipotezę dotyczącą superinteligencji: że sztuczna inteligencja wystarczająco inteligentna, aby zrozumieć swój własny projekt, mogłaby się przeprojektować lub stworzyć kolejny, bardziej inteligentny system, który mógłby następnie ponownie się przeprojektować, aby stać się jeszcze bardziej inteligentny, i tak włączone w cyklu pozytywnego sprzężenia zwrotnego.

Good nazwał to „eksplozją inteligencji”. Scenariusze rekurencyjne nie ograniczają się do sztucznej inteligencji: ludzie, których inteligencja jest wzmocniona poprzez interfejs mózg-komputer, mogą zwrócić uwagę na projektowanie interfejsów mózg-komputer nowej generacji. (Gdybyś miał maszynę, która zwiększyła twoje IQ, z pewnością przyszłoby ci do głowy, gdy staniesz się wystarczająco mądry, aby spróbować zaprojektować potężniejszą wersję maszyny.) Superinteligencję można również osiągnąć poprzez zwiększenie szybkości przetwarzania. Najszybsze zaobserwowane neurony uruchamiają się 1000 razy na sekundę; najszybsze włókna aksonów przewodzą sygnały z prędkością 150 metrów na sekundę, co stanowi półmilionową prędkość światła (Sandberg 1999). Wydaje się, że powinno być fizycznie możliwe zbudowanie mózgu, który wykonuje obliczenia milion razy szybciej niż mózg ludzki, bez zmniejszania jego rozmiarów i przepisywania oprogramowania. Gdyby ludzki umysł został w ten sposób przyspieszony, subiektywny rok myślenia przypadałby na każde trzydzieści jeden fizycznych sekund w świecie zewnętrznym, a tysiąclecie minęłoby w osiem i pół godziny. Vinge (1993) nazwał takie przyspieszone umysły „słabą superinteligencją”: umysłem, który myśli jak człowiek, ale znacznie szybciej. Yudkowsky (2008a) wymienia trzy rodziny metafor służących do wizualizacji możliwości sztucznej inteligencji mądrzejszej od człowieka:

* Metafory inspirowane różnicami w indywidualnej inteligencji ludzi: sztuczna inteligencja będzie patentować nowe wynalazki, publikować przełomowe artykuły badawcze, zarabiać na giełdzie lub przewodzić blokom władzy politycznej.

* Metafory inspirowane różnicami w wiedzy między przeszłymi i obecnymi cywilizacjami ludzkimi: Szybka sztuczna inteligencja wynajdzie możliwości, które futuryści powszechnie przewidują dla cywilizacji ludzkich za sto lub tysiąclecie przyszłości, takie jak nanotechnologia molekularna czy podróże międzygwiezdne.

* Metafory inspirowane różnicami w architekturze mózgu człowieka i innych organizmów biologicznych: Na przykład: „Wyobraźcie sobie, że umysł psa działa z bardzo dużą szybkością. Czy tysiąc lat życia na psie przyniosłoby jakikolwiek ludzki wgląd?” (Vinge’a (1993)). Oznacza to, że zmiany w architekturze poznawczej mogą generować spostrzeżenia, których żaden umysł na poziomie ludzkim nie byłby w stanie znaleźć ani być może nawet przedstawić po jakimkolwiek czasie.

Nawet jeśli ograniczymy się do metafor historycznych, stanie się jasne, że nadludzka inteligencja stwarza wyzwania etyczne, które są dosłownie bezprecedensowe. W tym momencie stawka nie dotyczy już skali indywidualnej (np. bezpodstawna odmowa kredytu hipotecznego, pożar domu, złe traktowanie osoby-agenta), ale globalna lub kosmiczna (np. ludzkość zostaje wygaśnięta i zastąpiona przez nic, co uznalibyśmy za wartościowe).). Lub, jeśli superinteligencję można ukształtować tak, aby była korzystna, wówczas, w zależności od jej możliwości technologicznych, może szybko rozwiązać wiele współczesnych problemów, które okazały się trudne dla naszej inteligencji na poziomie ludzkim. Superinteligencja jest jednym z kilku „ryzyk egzystencjalnych” zgodnie z definicją Bostroma (2002): ryzyko, „w przypadku którego niekorzystny wynik albo unicestwi inteligentne życie pochodzące z Ziemi, albo trwale i drastycznie ograniczy jego potencjał”. I odwrotnie, pozytywny wynik superinteligencji mógłby chronić inteligentne życie pochodzące z Ziemi i pomóc w wykorzystaniu jego potencjału. Należy podkreślić, że mądrzejsze umysły niosą ze sobą ogromne potencjalne korzyści, ale także ryzyko. Próby uzasadnienia globalnego ryzyka katastroficznego mogą być podatne na szereg błędów poznawczych , w tym na „błąd dobrej historii” zaproponowany przez Bostroma (2002):

Załóżmy, że nasze intuicje dotyczące tego, które przyszłe scenariusze są „prawdopodobne i realistyczne”, kształtują się na podstawie tego, co widzimy w telewizji i filmach oraz co czytamy w powieściach. (W końcu duża część dyskursu o przyszłości, z którą spotykają się ludzie, ma formę fikcji i innych kontekstów rekreacyjnych.) Powinniśmy zatem, myśląc krytycznie, podejrzewać, że nasza

intuicja jest stronnicza w kierunku przeceniania prawdopodobieństwa te scenariusze, które składają się na dobrą historię, ponieważ takie scenariusze będą wydawać się znacznie bardziej znajome i bardziej „prawdziwe". To nastawienie do Dobrej Historii może być dość potężne. Kiedy ostatni raz widziałeś film o nagłym wyginięciu ludzkości (bez ostrzeżenia i bez zastąpienia jej inną cywilizacją)? Chociaż ten scenariusz może być znacznie bardziej prawdopodobny niż scenariusz, w którym ludzcy bohaterowie skutecznie odpierają inwazję potworów lub robotycznych wojowników, oglądanie go nie byłoby zbyt przyjemne.

Naprawdę pożądane rezultaty sprawiają, że filmy są kiepskie – brak konfliktu oznacza brak historii. Chociaż Trzy prawa robotyki Asimova (Asimov 1942) są czasami cytowane jako model etycznego rozwoju sztucznej inteligencji, Trzy prawa są w równym stopniu narzędziem fabularnym, jak „pozytronowy mózg" Asimova. Gdyby Asimov przedstawił Trzy Prawa jako dobrze działające, nie miałby żadnych historii. Błędem byłoby uważać „AI" za gatunek o ustalonych cechach i pytać: „Czy będą dobre, czy złe?" Termin „sztuczna inteligencja" odnosi się do ogromnej przestrzeni projektowej, prawdopodobnie znacznie większej niż przestrzeń ludzkich umysłów (ponieważ wszyscy ludzie mają wspólną architekturę mózgu). Pytanie: „Czy sztuczna inteligencja będzie dobra, czy zła?" może być formą nastawienia na dobrą historię. jakby próbował wybrać przesłankę do fabuły filmu. Odpowiedź powinna brzmieć: „O którym dokładnie projekcie sztucznej inteligencji mówisz?" Czy kontrola nad wstępnym zaprogramowaniem sztucznej inteligencji może przełożyć się na jej późniejszy wpływ na świat? Kurzweil (2005) utrzymuje, że „inteligencja jest z natury niemożliwa do kontrolowania" i że pomimo wszelkich prób podejmowanych przez człowieka środków ostrożności, „z definicji… inteligentne istoty są na tyle sprytne, aby łatwo pokonać takie bariery". Załóżmy, że sztuczna inteligencja jest nie tylko sprytna, ale w ramach procesu doskonalenia własnej inteligencji ma nieograniczony dostęp do własnego kodu źródłowego – może przepisać się na wszystko, czym chce. Nie oznacza to jednak, że sztuczna inteligencja musi chcieć przepisać się do wrogiej formy. Weźmy pod uwagę Gandhiego, który najwyraźniej szczerze pragnął nie zabijać ludzi. Gandhi nie wziąłby świadomie pigułki, która spowodowała, że chce zabijać ludzi, ponieważ Gandhi wie, że jeśli chce zabijać ludzi, prawdopodobnie ich zabije, a obecna wersja Gandhiego nie chce zabijać. Mówiąc bardziej ogólnie, wydaje się prawdopodobne, że większość samomodyfikujących się umysłów będzie w naturalny sposób miała stabilne funkcje użyteczności, co oznacza, że początkowy wybór projektu umysłu może mieć trwałe skutki (Omohundro 2008). Czy na tym etapie rozwoju nauki o sztucznej inteligencji można w jakiś sposób przełożyć zadanie znalezienia projektu „dobrej" sztucznej inteligencji na nowoczesny kierunek badań? Spekulacje mogą wydawać się przedwczesne, ale można podejrzewać, że niektóre paradygmaty sztucznej inteligencji z większym prawdopodobieństwem niż inne ostatecznie okażą się pomocne w stworzeniu inteligentnych czynników samomodyfikujących się, których cele pozostają przewidywalne nawet po wielokrotnych iteracjach samodoskonalenia. Na przykład Bayesowska gałąź sztucznej inteligencji, inspirowana spójnymi systemami matematycznymi, takimi jak teoria prawdopodobieństwa i maksymalizacja oczekiwanej użyteczności, wydaje się bardziej podatna na przewidywalny problem samomodyfikacji niż programowanie ewolucyjne i algorytmy genetyczne. Jest to kontrowersyjne stwierdzenie, ale ilustruje fakt, że jeśli pomyślimy o wyzwaniach, jakie w przyszłości stanowić będzie superinteligencja, rzeczywiście można to przekształcić w wskazówki dotyczące kierunków obecnych badań nad sztuczną inteligencją. Jednak nawet zakładając, że możemy określić system celów sztucznej inteligencji, który będzie trwały w przypadku samomodyfikacji i samodoskonalenia, zaczyna to jedynie dotykać podstawowych problemów etycznych związanych z tworzeniem superinteligencji. Ludzie, pierwsza inteligencja ogólna istniejąca na Ziemi, wykorzystała tę inteligencję do zasadniczego przekształcenia globu – rzeźbienia gór, oswajania rzek, budowania drapaczy chmur, uprawiania pustyń, powodowania niezamierzonych planetarnych zmian klimatycznych. Silniejsza inteligencja może mieć odpowiednio większe konsekwencje. Rozważmy

ponownie historyczną metaforę superinteligencji – różnice podobne do różnic między cywilizacjami przeszłymi i obecnymi. Naszą obecną cywilizację nie oddziela od starożytnej Grecji jedynie udoskonalona nauka i zwiększone możliwości technologiczne. Istnieje różnica punktów widzenia etycznego: starożytni Grecy uważali, że niewolnictwo jest dopuszczalne; myślimy inaczej. Nawet między XIX a XX wiekiem istniały istotne rozbieżności etyczne – czy kobiety powinny mieć prawo głosu? Czy czarni powinni mieć głos? Wydaje się prawdopodobne, że przyszłe cywilizacje nie będą postrzegać dzisiejszych ludzi jako etycznie doskonałych – nie tylko z powodu tego, że nie rozwiązaliśmy obecnie uznawanych problemów etycznych, takich jak ubóstwo i nierówność, ale także dlatego, że nawet nie rozpoznaliśmy pewnych problemów etycznych. Być może pewnego dnia poddawanie dzieci przymusowej nauce będzie postrzegane jako znęcanie się nad dziećmi, a może pozwolenie dzieciom na opuszczenie szkoły w wieku 18 lat będzie postrzegane jako znęcanie się nad dziećmi. Nie wiemy.

Biorąc pod uwagę etyczną historię cywilizacji ludzkich na przestrzeni wieków, widzimy, że wielką tragedią mogłoby okazać się stworzenie umysłu stabilnego pod względem etycznym, w którym cywilizacje ludzkie wydają się wykazywać zmiany kierunkowe. Co by było, gdyby Archimedes z Syrakuz był w stanie stworzyć długotrwały sztuczny intelekt z ustaloną wersją kodeksu moralnego starożytnej Grecji? Jednak uniknięcie tego rodzaju etycznej stagnacji może okazać się trudne. Nie wystarczyłoby na przykład po prostu uczynić umysł losowo niestabilnym. Starożytni Grecy, nawet gdyby zdawali sobie sprawę ze swojej niedoskonałości, nie mogliby zrobić lepiej, rzucając kostkami. Czasami pojawia się nowy, dobry pomysł z zakresu etyki, który jest zaskoczeniem; ale większość losowo generowanych zmian etycznych wydałaby nam się szaleństwem lub bełkotem. Stawia to nas przed być może ostatecznym wyzwaniem związanym z etyką maszyn: jak zbudować sztuczną inteligencję, która po uruchomieniu stanie się bardziej etyczna od Ciebie? To nie jest tak, jakby prosić naszych filozofów o stworzenie superetyki, tak samo jak Deep Blue nie został stworzony poprzez nakłonienie najlepszych szachistów do programowania dobrych ruchów. Musimy jednak umieć skutecznie opisać pytanie, jeśli nie odpowiedź – rzucanie kostkami nie zapewni dobrych ruchów szachowych ani dobrej etyki. Lub być może bardziej produktywny sposób myślenia o problemie: jaką strategię chciałbyś, aby Archimedes zastosował przy budowaniu superinteligencji, tak aby ogólny wynik był nadal akceptowalny, gdybyś nie mógł mu powiedzieć, co konkretnie robi źle? To jest bardzo podobna sytuacja w której się znajdujemy, w odniesieniu do przyszłości. Jedna mocna rada, która wyłania się z uznania naszej sytuacji za analogiczną do sytuacji Archimedesa, jest taka, że nie powinniśmy próbować wymyślać „super" wersji tego, co nasza cywilizacja uważa za etykę – nie jest to strategia, której byśmy chcieli Archimedes podążać. Być może powinniśmy raczej rozważyć pytanie, w jaki sposób sztuczna inteligencja zaprogramowana przez Archimedesa, nie mająca większej wiedzy moralnej niż Archimedes, mogłaby rozpoznać (przynajmniej część) etyki naszej własnej cywilizacji jako postęp moralny, a nie zwykłą niestabilność moralną. Wymagałoby to, abyśmy zaczęli pojmować strukturę kwestii etycznych w taki sposób, w jaki pojmowaliśmy już strukturę szachów. Jeśli poważnie myślimy o rozwoju zaawansowanej sztucznej inteligencji, jest to wyzwanie, któremu musimy sprostać. Jeśli maszyny mają zostać umieszczone w pozycji silniejszej, szybszej, budzącej większe zaufanie lub mądrzejsze od ludzi, wówczas dyscyplina etyki maszyn musi zobowiązać się do poszukiwania wyższej niż ludzka (a nie tylko ludzkiej) uprzejmości.

Wniosek

Chociaż obecna sztuczna inteligencja stwarza nam niewiele problemów etycznych, które nie występują już w projektowaniu samochodów czy elektrowni, podejście algorytmów sztucznej inteligencji do myślenia bardziej ludzkiego zwiastuje przewidywalne komplikacje. Role społeczne mogą pełnić algorytmy sztucznej inteligencji, co oznacza nowe wymagania projektowe, takie jak przejrzystość i przewidywalność. Wystarczająco ogólne algorytmy sztucznej inteligencji mogą nie działać już w

przewidywalnych kontekstach, co wymaga nowych rodzajów zapewnienia bezpieczeństwa i inżynierii sztucznych względów etycznych. Sztuczne inteligencje z wystarczająco zaawansowanymi stanami umysłowymi lub właściwymi rodzajami stanów będą miały status moralny, a niektóre z nich mogą być liczone jako osoby – choć być może osoby bardzo różniące się od tych, które istnieją obecnie, być może rządzące się innymi zasadami. I wreszcie perspektywa sztucznej inteligencji dysponującej nadludzką inteligencją i nadludzkimi zdolnościami stawia nas przed niezwykłym wyzwaniem, jakim jest opracowanie algorytmu generującego nadetyczne zachowanie. Wyzwania te mogą wydawać się wizjonerskie, jednak można przewidzieć, że je napotkamy, i nie są one pozbawione sugestii dla współczesnych kierunków badań.