O co toczy się gra

Wiele może pójść nie tak, gdy ślepo wierzymy w big data - CATHY O'NEIL, TED TALK, 2017

23 marca 2016 r. Microsoft wypuścił Taya, zaprojektowanego jako ekscytujący i nowy chatbot, który nie był tworzony ręcznie z góry, jak oryginalny chatbot Eliza, ale został opracowany w dużej mierze dzięki uczeniu się na podstawie interakcji użytkownika. Wcześniejszy projekt, Xiaoice, który prowadzi rozmowy w języku chińskim, odniósł ogromny sukces w Chinach, a Microsoft miał duże nadzieje. Niecały dzień później projekt został odwołany. Nieprzyjemna grupa użytkowników próbowała utopić Tay w rasistowskiej, seksistowskiej i antysemickiej nienawiści. Ohydna mowa wkracza, ohydna mowa wychodzi; biedny Tay publikował tweety w stylu „Kurwa nienawidzę feministek" i „Hitler miał rację: nienawidzę Żydów". Gdzie indziej w Internecie pojawiają się różnego rodzaju problemy, wielkie i małe. Można przeczytać o Alexach, które straszyły swoich właścicieli przypadkowym chichotem, systemami rozpoznawania twarzy iPhone′a, które myliły matkę i syna, oraz o Poopocalypse, bałaganie podobnym do Jacksona Pollacka, który zdarzył się więcej niż jeden raz, gdy Roomba zderzyła się z psimi odchodami. Mówiąc poważniej, istnieją wykrywacze mowy nienawiści, które można łatwo oszukać, systemy kandydatów do pracy, które utrwalają uprzedzenia, oraz przeglądarki internetowe i silniki rekomendacji oparte na narzędziach sztucznej inteligencji, które zostały oszukane, by popychać ludzi w kierunku absurdalnych teorii spiskowych. W Chinach system rozpoznawania twarzy używany przez policję wysłał bilet na spacer do niewinnej osoby, która akurat była znanym przedsiębiorcą, gdy zobaczyła jej zdjęcie z boku autobusu, nie zdając sobie sprawy z tego, że zdjęcie większe niż w rzeczywistości to nie to samo, co sam przedsiębiorca. Tesla, najwyraźniej w trybie „Przywołania", rozbił się podczas wycofywania się z garażu swoich właścicieli. Niejednokrotnie kosiarki zautomatyzowane okaleczyły lub zabiły jeże. Sztucznej inteligencji, którą mamy teraz, po prostu nie można ufać. Chociaż często robi to właściwie, nigdy nie wiemy, kiedy zaskoczy nas błędami, które są bezsensowne, a nawet niebezpieczne. A im więcej autorytetu im dajemy, tym bardziej powinniśmy się martwić. Niektóre usterki są łagodne, jak Alexa, która przypadkowo chichocze (lub budzi cię w środku nocy) lub iPhone, który automatycznie poprawia to, co miało oznaczać „Wszystkiego najlepszego, drogi Teodorze" na „Wszystkiego najlepszego, martwy Teodor." Ale inne - takie jak algorytmy promujące fałszywe wiadomości lub uprzedzenia wobec kandydatów do pracy - mogą stanowić poważne problemy. Raport grupy AI Now szczegółowo opisuje wiele takich problemów w systemach AI w wielu różnych zastosowaniach, w tym w ustalaniu uprawnień do Medicaid, skazaniu na karę pozbawienia wolności i ocenie nauczycieli. Błyskawiczne awarie na Wall Street spowodowały tymczasowe spadki na giełdzie i miały miejsce przerażające naruszenia prywatności (jak wtedy, gdy Alexa nagrała rozmowę i nieumyślnie wysłała ją do przypadkowej osoby z listy kontaktów właściciela); i wiele wypadków samochodowych, niektóre śmiertelne. Nie bylibyśmy zaskoczeni, gdyby w sieci elektrycznej pojawiła się poważna awaria wywołana sztuczną inteligencją. Jeśli zdarzy się to w upalne lato lub w środku zimy, może umrzeć duża liczba osób.

Nie oznacza to, że nie powinniśmy nie spać po nocach, martwiąc się o świat podobny do Skynetu, w którym roboty walczą z ludźmi, przynajmniej w przewidywalnej przyszłości. Roboty nie mają jeszcze inteligencji ani zręczności manualnej, aby niezawodnie poruszać się po świecie, z wyjątkiem starannie kontrolowanych środowisk. Ponieważ ich zdolności poznawcze są tak wąskie i ograniczone, nie ma końca sposobom ograniczania ich. Co ważniejsze, nie ma powodu sądzić, że roboty, w stylu science fiction, powstaną przeciwko nam. Po sześćdziesięciu latach sztucznej inteligencji nie ma najmniejszego śladu złośliwości; maszyny wykazywały zerowe zainteresowanie wiązaniem się z ludźmi w celu zdobycia terytorium, posiadłości, praw do przechwalania się lub czegokolwiek innego, o co toczono bitwy. Nie są przepełnione testosteronem ani niepohamowaną żądzą dominacji nad światem. Zamiast tego AI są nerdami i idiotami, uczonymi skupionymi tak mocno na tym, co robią, że nie są świadomi

większego obraz. Weźmy grę Go, grę, która nominalnie obraca się wokół zajmowania terytorium, co jest tak bliskie przejęcia świata, jak każda istniejąca sztuczna inteligencja. W latach 70. komputerowe programy Go były okropne, łatwo pokonane przez każdego przyzwoitego gracza, ale nie wykazywały żadnych oznak chęci ingerowania w ludzkość. Czterdzieści lat później programy takie jak AlphaGo są fantastyczne, znacznie ulepszone i znacznie lepsze niż najlepsi ludzie; ale nadal wykazują zerowe zainteresowanie przejmowaniem ludzkiego terytorium lub degradacją ich programistów i do zoo. Jeśli nie ma tego na tablicy, nie są zainteresowane. AlphaGo po prostu nie przejmuje się pytaniami typu „Czy istnieje życie poza tablicą Go?", nie mówiąc już o „Czy to sprawiedliwe, że moi ludzcy mistrzowie nie robią nic poza graniem w Go przez cały dzień?" AlphaGo dosłownie nie ma żadnego życia ani ciekawości poza planszą; nie wie, że gra jest zwykle rozgrywana za pomocą kamieni, ani nawet, że coś istnieje poza siatką, na której gra. Nie wie, że jest komputerem i korzysta z energii elektrycznej ani że jego przeciwnikiem jest człowiek. Nie pamięta, że grał w wiele gier w przeszłości, ani nie przewiduje, że będzie grał więcej gier w przyszłości. Nie jest zadowolony, gdy wygrywa, nie jest zmartwiony, gdy przegrywa, ani nie jest dumny z postępów, jakie poczynił w nauce gry w Go. Rodzaje ludzkich motywów, które napędzają agresję w świecie rzeczywistym, są całkowicie nieobecne. Gdybyś chciał spersonalizować algorytm (jeśli to w ogóle ma sens), powiedziałbyś, że AlphaGo jest całkowicie zadowolony z tego, co robi, bez chęci robienia czegokolwiek innego. To samo można powiedzieć o sztucznej inteligencji, która zajmuje się diagnostyką medyczną, zaleceniami reklamowymi, nawigacją lub czymkolwiek innym. Maszyny, przynajmniej w ich obecnej implementacji, pracują nad zadaniami, do których zostały zaprogramowane, i niczym więcej. Dopóki zachowujemy się w ten sposób, nasze zmartwienia nie powinny skupiać się na jakiejś wyimaginowanej złośliwości. Jak napisał Steven Pinker:

Scenariusz [że roboty staną się superinteligentnymi i zniewolą ludzi] ma tyle samo sensu, co obawa, że skoro samoloty odrzutowe przewyższają zdolność latania orłów, pewnego dnia spadną z nieba i schwytają nasze bydło. Błąd … to pomieszanie inteligencji z motywacją - przekonań z pragnieniami, wnioskowania z celami, myślenia z pragnieniem. Nawet gdybyśmy wynaleźli nadludzko inteligentne roboty, dlaczego mieliby chcieć zniewolić swoich panów lub przejąć władzę nad światem? Inteligencja to zdolność do wdrażania nowatorskich środków do osiągnięcia celu. Ale cele są obce inteligencji: bycie mądrym to nie to samo, co chcieć czegoś.

Aby przejąć świat, roboty musiałyby chcieć; musiałyby być agresywne, ambitne i niezadowolone, z brutalną passą. Takiego robota jeszcze nie spotkaliśmy. A na razie nie ma żadnego powodu, by budować roboty ze stanami emocjonalnymi, ani przekonującego pomysłu na to, jak moglibyśmy to zrobić, nawet gdybyśmy chcieli. Ludzie mogą używać emocji, takich jak niezadowolenie, jako narzędzia motywacji, ale roboty nie potrzebują niczego, aby pojawić się w pracy; po prostu robią to, co im każą. Nie wątpimy, że roboty mogą pewnego dnia mieć fizyczne i intelektualne moce, które mogą uczynić z nich groźnych wrogów - jeśli zechcą nam się przeciwstawić - ale przynajmniej w przewidywalnej przyszłości nie widzimy żadnego powodu, dla którego by to zrobiły.

Mimo to nie jesteśmy wolni w domu. Sztuczna inteligencja nie musi chcieć nas zniszczyć, aby wywołać spustoszenie. W krótkiej perspektywie najbardziej powinniśmy się martwić, czy maszyny rzeczywiście są w stanie niezawodnie wykonywać zadania, które im zlecamy. Cyfrowy asystent, który ustala terminy naszych wizyt, jest nieoceniony – jeśli jest niezawodny. Jeśli przypadkowo wyśle nas na krytyczne spotkanie z tygodniowym opóźnieniem, to katastrofa. A jeszcze więcej będzie zagrożone, gdy dokonamy nieuniknionego przejścia na roboty domowe. Jeśli jakiś korporacyjny tytan projektuje domowego robota do robienia crème brûlée, chcemy, żeby działał za każdym razem, a nie tylko dziewięć razy na dziesięć, podpalając kuchnię po raz dziesiąty. Maszyny, o ile nam wiadomo, nigdy nie mają imperialistycznych ambicji, ale popełniają błędy, a im bardziej na nich polegamy, tym większe znaczenie mają ich błędy. Innym problemem, na razie całkowicie nierozwiązanym, jest to, że maszyny

muszą poprawnie wnioskować o naszych intencjach, nawet jeśli daleko nam do jednoznacznych, a nawet w oczywisty sposób niejasnych. Jedną z kwestii jest to, co moglibyśmy nazwać problemem Amelii Bedeli, na cześć gospodyni w serii opowiadań dla dzieci, która zbyt dosłownie traktuje prośby swojego pracodawcy. Wyobraź sobie, że rano mówisz swojemu robotowi sprzątającemu: „Weź wszystko, co zostało w salonie i schowaj do szafy", tylko po to, aby wrócić i zobaczyć wszystko – telewizor, meble, dywan – rozbite małe kawałki do dopasowania. Następnie pojawia się problem błędów mowy, szczególnie prawdopodobnych w opiece nad starszymi z trudnościami poznawczymi. Jeśli dziadek prosi, aby obiad wyrzucić do śmieci, a nie na stół, dobry robot powinien mieć rozsądek, aby upewnić się, że tego naprawdę chce, a nie pomyłka. Posługując się ostatnio spopularyzowaną frazą, chcemy, aby nasze roboty i sztuczna inteligencja traktowały nas poważnie, ale nie zawsze dosłownie.

Oczywiście każda technologia może zawieść, nawet najstarsza i najlepiej zrozumiana. Niedługo przed rozpoczęciem pracy nad tą książką .Chodnik dla pieszych w Miami zawalił się spontanicznie, zabijając sześć osób, zaledwie pięć dni po jego zainstalowaniu, mimo że ludzie budowali mosty od ponad trzech tysiącleci (most Arkadiko, zbudowany w 1300 r. p.n.e., nadal stoi). Nie możemy oczekiwać, że sztuczna inteligencja będzie idealna od pierwszego dnia, a w niektórych przypadkach istnieje dobry argument za tolerowaniem krótkoterminowego ryzyka w celu osiągnięcia długoterminowych korzyści; gdyby teraz kilka osób zginęło podczas opracowywania samochodów bez kierowcy, ale ostatecznie uratowano setki tysięcy lub miliony, ryzyko może być warte podjęcia. To powiedziawszy, dopóki sztuczna inteligencja nie zostanie przeformułowana i ulepszona w fundamentalny sposób, ryzyko jest obfite. Oto dziewięć wyzwań, o które martwimy się najbardziej. Po pierwsze, mamy do czynienia z podstawowym błędem nadmiernej atrybucji. Sztuczna inteligencja często skłania nas do uwierzenia, że ma inteligencję podobną do ludzkiej, nawet jeśli jej nie ma. Jak zauważyła autorka i socjolog z MIT, Sherry Turkle, przyjazny pozornie robot towarzyszący nie jest w rzeczywistości twoim przyjacielem. I możemy zbyt szybko scedować władzę na sztuczną inteligencję, zakładając, że sukces w jednym kontekście zapewnia niezawodność w innym. Jeden z najbardziej oczywistych przykładów, o których już wspomnieliśmy: przypadek samochodów bez kierowcy - dobre osiągi w zwykłych warunkach nie gwarantują bezpieczeństwa we wszystkich okolicznościach. Weźmy bardziej subtelny przykład: niedawno policjanci w Kansas zatrzymali kierowcę i użyli Tłumacza Google, aby uzyskać zgodę kierowcy na przeszukanie jego samochodu. Sędzia stwierdził później, że jakość tłumaczenia była tak słaba, że kierowca nie mógł być uznany za osobę świadomą zgody, a zatem uznał, że przeszukanie jest sprzeczne z Czwartą Poprawką. Dopóki sztuczna inteligencja nie ulegnie radykalnemu polepszeniu, musimy być ostrożni, by za bardzo jej ufać. Po drugie, brakuje solidności. Jednym z przykładów jest potrzeba, aby samochody bez kierowców radziły sobie z nietypowym oświetleniem, nietypową pogodą, nietypowymi śmieciami na drodze, nietypowymi wzorcami ruchu, ludźmi wykonującymi nietypowe gesty i tak dalej. Podobnie solidność byłaby niezbędna dla systemu, który naprawdę przejąłby kontrolę nad Twoim kalendarzem; jeśli się zdezorientuje podczas podróży z Kalifornii do Bostonu i spóźniłeś się trzy godziny na spotkanie, masz problem. Potrzeba lepszego podejścia do sztucznej inteligencji jest oczywista. Po trzecie, współczesne uczenie maszynowe w dużej mierze zależy od dokładnych szczegółów dużych zestawów szkoleniowych, a takie systemy często ulegają awarii, jeśli są stosowane do nowych problemów, które wykraczają poza konkretne zestawy danych, na których zostały przeszkolone. Systemy tłumaczenia maszynowego przeszkolone w zakresie dokumentów prawnych radzą sobie słabo, gdy są stosowane do artykułów medycznych i na odwrót. Systemy rozpoznawania głosu nauczone tylko dla dorosłych native speakerów często mają problemy z akcentem. Technologia, podobnie jak to, na czym opiera się Tay, działała dobrze, gdy wzięła swój wkład od społeczeństwa, w którym mowa polityczna jest mocno uregulowana, ale dała niedopuszczalne wyniki, gdy utonęła w morzu inwektyw. System głębokiego uczenia, który potrafił rozpoznawać cyfry wydrukowane na

czarno na białym tle z 99-procentową dokładnością, znikał, gdy kolory zostały odwrócone, nagle uzyskując tylko 34 procent poprawnej wartości - mało pocieszające, gdy przestajesz zastanawiać się nad tym, że są niebieskie stop znaki na Hawajach. Informatyk ze Stanford, Judy Hoffman, wykazała, że autonomiczny pojazd, którego system wizyjny został wyszkolony w jednym mieście, może radzić sobie znacznie gorzej w innym, nawet jeśli chodzi o rozpoznawanie podstawowych obiektów, takich jak drogi, znaki drogowe i inne samochody. Po czwarte, pogłębianie danych na ślepo może prowadzić do utrwalania przestarzałych uprzedzeń społecznych, gdy potrzebne jest bardziej subtelne podejście. Jedna z pierwszych oznak tego zjawiska pojawiła się w 2013 roku, kiedy Latanya Sweeney, informatyk z Harvardu, odkrył, że po wyszukaniu w Google tak charakterystycznie czarnego nazwiska, jak Jermaine, otrzymywało się znacznie więcej reklam zawierających informacje o aktach aresztowania niż w przypadku wyszukiwania hasła charakterystycznie białe imię, takie jak Geoffrey. Następnie, w 2015 r., zdjęcia Google błędnie oznaczyły niektóre zdjęcia Afroamerykanów jako goryle. W 2016 roku ktoś odkrył, że jeśli wyszukasz w Google hasło „Profesjonalna fryzura do pracy", zwrócone zdjęcia przedstawiały prawie wszystkie białe kobiety, podczas gdy w wyszukiwarce „Nieprofesjonalna fryzura do pracy" były to prawie wszystkie czarne kobiety. W 2018 roku Joy Buolamwini, wówczas absolwentka MIT Media Lab, odkryła, że wiele komercyjnych algorytmów ma tendencję do błędnej identyfikacji płci Afroamerykanek. IBM jako pierwszy załatał ten konkretny problem, a Microsoft szybko poszedł w jego ślady - ale o ile wiemy, nikt nie wymyślił rozwiązania ogólnego. Nawet teraz, kiedy to piszemy, łatwo znaleźć podobne przykłady. Kiedy wyszukaliśmy w Google hasło „matka", przeważająca większość obrazów przedstawiały białych ludzi, artefakt gromadzenia danych w sieci i oczywiste zniekształcenie rzeczywistości. Kiedy szukaliśmy hasła „profesor", tylko około 10 procent najwyżej ocenianych obrazów to kobiety, być może odzwierciedlające hollywoodzki styl życia na studiach, ale nie mające kontaktu z obecną rzeczywistością, w której blisko połowa profesorów to kobiety. System rekrutacji oparty na sztucznej inteligencji, który Amazon uruchomił w 2014 r., był tak problematyczny, że musieli go porzucić w 2018 r. Nie uważamy, że te problemy są nie do pokonania - jak omówimy później, zmiana paradygmatu w sztucznej inteligencji może tu pomóc - ale nie ma ogólnego rozwiązania jeszcze istnieje. Podstawową kwestią jest to, że obecne systemy sztucznej inteligencji naśladują dane wejściowe, bez względu na wartości społeczne, jakość czy charakter danych. Statystyki rządu USA mówią nam, że obecnie tylko 41 procent wykładowców to biali mężczyźni, ale wyszukiwarka grafiki Google o tym nie wie; po prostu miesza każdy znaleziony obraz, nie zastanawiając się nad jakością i reprezentatywnością danych lub wartości, które są wyrażane w sposób dorozumiany. Dane demograficzne wydziałów zmieniają się, ale ślepe pogłębiarki danych tęsknią za tym i raczej umacniają historię niż odzwierciedlają zmieniającą się rzeczywistość. Podobne obawy pojawiają się, gdy myślimy o roli, jaką AI zaczyna odgrywać w medycynie. Zestawy danych wykorzystywane do szkolenia programów diagnostycznych raka skóry, na przykład, mogą być dostosowane do pacjentów rasy białej i dają nieprawidłowe wyniki, gdy są stosowane u pacjentów z ciemniejszą karnacją. Autonomiczne samochody mogą być wymiernie mniej niezawodne w rozpoznawaniu ciemnoskórych pieszych niż jasnoskórych. Stawką jest życie, a obecne systemy nie są przygotowane aby zmagać się z tymi uprzedzeniami. Po piąte, silna zależność współczesnej sztucznej inteligencji od zestawów treningowych może również prowadzić do szkodliwego efektu komory echa, w którym system jest trenowany na danych, które sam wygenerował wcześniej. Na przykład, jak omówimy później, programy tłumaczeniowe działają poprzez uczenie się z „bittekstów", pary dokumentów, które są wzajemnie tłumaczeniami. Niestety istnieją języki, w których znaczna część tekstów w sieci - w niektórych przypadkach nawet 50 procent wszystkich dokumentów internetowych - została w rzeczywistości stworzona przez program do tłumaczenia maszynowego. W rezultacie, jeśli Tłumacz Google popełni jakiś błąd w tłumaczeniu, ten błąd może skończyć się w dokumencie w Internecie, a dokument ten staje się danymi, wzmacniając błąd. Podobnie, wiele systemów polega na ludzkich pracownikach społecznościowych do oznaczania obrazów, ale czasami pracownicy społecznościowi używają botów wykorzystujących sztuczną inteligencję, aby wykonywać

swoją pracę za nich. Chociaż społeczność badawcza AI opracowała z kolei techniki sprawdzania, czy praca jest wykonywana przez ludzi czy boty, cały proces stał się grą w kotka i myszkę między badaczami AI z jednej strony i psotnymi botami tłumu z drugiej , przy czym żadna ze stron nie utrzymuje trwałej przewagi. W rezultacie wiele rzekomo wysokiej jakości danych oznaczonych przez ludzi okazuje się być generowanych maszynowo. Po szóste, programy, które opierają się na danych, którymi może manipulować opinia publiczna, są często podatne na oszukiwanie. Tay jest oczywiście tego przykładem. A w Google regularnie trafiają „bomby Google", w których ludzie tworzą dużą liczbę postów i linków, dzięki czemu wyszukiwanie określonych haseł daje wyniki, które uważają za zabawne. Na przykład w lipcu 2018 r. ludziom udało się nakłonić Grafika Google do odpowiedzi na wyszukiwania hasła „idiota" ze zdjęciami Donalda Trumpa. (Było to nadal prawdą w tym samym roku, kiedy Sundar Pichai przemawiał w Kongresie). Szesnaście lat wcześniej miała miejsce kolejna parodia, raczej bardziej nieprzyzwoita, Ricka Santoruma. A ludzie nie tylko grają w Google dla chichotów; cała branża, optymalizacja pod kątem wyszukiwarek, polega na manipulowaniu Google w celu zapewnienia wysokiej pozycji swoim klientom w odpowiednich wyszukiwaniach internetowych. Po siódme, połączenie istniejących uprzedzeń społecznych i efektu echa może prowadzić do wzmocnienia uprzedzeń społecznych. Załóżmy, że w przeszłości w jakimś mieście policja, wyroki skazujące i wyroki były niesprawiedliwie stronnicze wobec określonej grupy mniejszościowej. Miasto decyduje się teraz na użycie programu Big Data, aby doradzać mu w sprawach policyjnych i skazujących, a program jest szkolony na danych historycznych, w których groźnych przestępców identyfikuje się pod względem ich historii aresztowań i czasu więzienia. Program dostrzeże, że tak zdefiniowani niebezpieczni przestępcy nieproporcjonalnie pochodzą z mniejszości; w związku z tym zaleci, aby dzielnice z większym odsetkiem tych mniejszości otrzymały więcej policji, a członkowie mniejszości powinni być szybciej aresztowani i skazywani na dłuższe wyroki. Gdy program zostanie ponownie uruchomiony na nowym zbiorze danych, nowe dane potwierdzają jego wcześniejsze osądy, a program będzie miał tendencję do formułowania tego samego rodzaju stronniczych zaleceń z jeszcze większą pewnością. Jak podkreśliła Cathy O'Neil, autorka Weapons of Math Destruction, nawet jeśli program jest napisany tak, aby uniknąć używania rasy i pochodzenia etnicznego jako kryteriów, istnieją wszelkiego rodzaju funkcje związane z „proxy", których mógłby zamiast tego użyć to miałoby ten sam wynik: sąsiedztwo, powiązania z mediami społecznościowymi, wykształcenie, praca, język, a może nawet takie rzeczy, jak preferencje w ubiorze. Co więcej, decyzje, które podejmuje program, są obliczane „algorytmicznie", mają aurę obiektywizmu, która imponuje biurokratom i dyrektorom firm oraz krowom opinii publicznej. Działanie programów jest tajemnicze - dane treningowe są poufne, program jest zastrzeżony, a proces podejmowania decyzji jest „czarną skrzynką", której nawet projektanci programu nie potrafią wyjaśnić - tak więc osoby fizyczne stają się prawie niemożliwe do zakwestionowania decyzji, które czują się niesprawiedliwe. Kilka lat temu Xerox chciał ograniczyć kosztowną „rezygnację" wśród pracowników, więc wdrożył program Big Data, aby przewidzieć, jak długo będzie trwał pracownik. Program wykazał, że jedną wysoce predykcyjną zmienną była długość dojazdów; nic dziwnego, że pracownicy z długimi dojazdami mają tendencję do wcześniejszego odchodzenia z pracy. Jednak kierownictwo Xerox zdało sobie sprawę, że niezatrudnianie osób z długimi dojazdami oznaczałoby w efekcie dyskryminację osób o niskich lub średnich dochodach, ponieważ firma była zlokalizowana w zamożnej dzielnicy. Na swoją korzyść firma usunęła to jako rozważane kryterium. Ale bez ścisłego monitorowania ludzi tego rodzaju uprzedzenia z pewnością będą się pojawiać. Ósmym wyzwaniem dla AI jest to, że AI zbyt łatwo może skończyć z niewłaściwymi celami. Badaczka DeepMind Victoria Krakovna zebrała dziesiątki przykładów tego zdarzenia. Robot grający w piłkę nożną, zachęcany do prób dotykania piłki tyle razy, ile się da, opracował strategię stania obok piłki i szybkiego wibrowania - nie do końca o to chodziło programistom. Robot, który miał nauczyć się chwytać konkretny przedmiot, był szkolony na obrazach tego, jak to wygląda, gdy chwyta ten przedmiot, więc uznał, że wystarczy włożyć rękę między aparat a przedmiot, aby wyglądał, jakby robot chwytał przedmiot. Niezbyt ambitna sztuczna inteligencja, której

zadaniem jest granie w Tetrisa, zdecydowała, że lepiej zatrzymać grę na czas nieokreślony, niż ryzykować przegraną. Problem niedopasowanych celów może również przybierać subtelniejsze formy. We wczesnych dniach uczenia maszynowego firma mleczarska wynajęła firmę zajmującą się uczeniem maszynowym do zbudowania systemu, który mógłby przewidywać, kiedy krowy wejdą w ruję. Określonym celem programu było jak najdokładniejsze generowanie prognoz „ruja/brak rui". Rolnicy byli zachwyceni, gdy dowiedzieli się, że system był dokładny w 95 procentach. Byli mniej zadowoleni, gdy dowiedzieli się, jak program tym zarządza. Krowy są w rui tylko przez jeden dzień w dwudziestodniowym cyklu; na tej podstawie przewidywania programu dla każdego dnia były takie same (bez rui), co sprawiało, że program był poprawny w dziewiętnastu dniach z dwudziestu – i całkowicie bezużyteczny. O ile nie opiszemy tego szczegółowo, rozwiązanie, którego chcemy, może nie być tym, na którym działa system AI. Wreszcie, ze względu na skalę, na jaką może działać obecna sztuczna inteligencja, istnieje wiele sposobów, na które sztuczna inteligencja może (nawet w swojej wciąż prymitywnej formie) być wykorzystywana celowo do wyrządzania poważnej szkody publicznej. Stalkerzy zaczęli używać stosunkowo podstawowych technik sztucznej inteligencji do monitorowania i manipulowania swoimi ofiarami, a spamerzy od lat wykorzystują sztuczną inteligencję do identyfikowania potencjalnych znaków, unikania CAPTCHA na stronach internetowych, które zapewniają, że jesteś człowiekiem i tak dalej. Nie ma wątpliwości, że sztuczna inteligencja wkrótce znajdzie rolę w autonomicznych systemach broni, choć mamy skromną nadzieję, że takie techniki mogą zostać zakazane, podobnie jak broń chemiczna. Jako politolog SUNY Virginia Eubanks zauważyła: „Kiedy bardzo skuteczna technologia zostanie zastosowana przeciwko pogardzanej grupie obcej przy braku silnej ochrony praw człowieka, istnieje ogromny potencjał okrucieństwa".

Nic z tego nie oznacza, że sztuczna inteligencja nie może działać lepiej – ale tylko wtedy, gdy nastąpi fundamentalna zmiana paradygmatu, taka, o jaką apelujemy. Jesteśmy przekonani, że wiele z tych problemów technicznych można rozwiązać - ale nie obecne techniki. To, że współczesna sztuczna inteligencja jest napędzana niewolniczo przez dane bez prawdziwego zrozumienia wartości etycznych, które programiści i projektanci systemów mogą chcieć przestrzegać, nie oznacza, że cała sztuczna inteligencja w przyszłości musi być podatna na te same problemy. Ludzie również patrzą na dane, ale nie zamierzamy decydować, że prawie wszyscy ojcowie i córki są biali lub że zadaniem piłkarza, który został zachęcony do częstszego dotykania piłki, jest stanie obok piłki i wibrowanie . Jeśli ludzie mogą uniknąć tych błędów, maszyny również powinny. Nie chodzi o to, że w zasadzie niemożliwe jest zbudowanie fizycznego urządzenia, które może jeździć po śniegu lub być etyczne; chodzi o to, że nie możemy się tam dostać samymi big data. To, czego naprawdę potrzebujemy, to zupełnie nowe podejście, z dużo większym wyrafinowaniem w kwestii tego, czego chcemy przede wszystkim: sprawiedliwego i bezpiecznego świata. Zamiast tego widzimy techniki sztucznej inteligencji, które rozwiązują indywidualne, wąskie problemy, jednocześnie omijając podstawowe problemy, które mają rozwiązać; mamy plastry, kiedy potrzebujemy przeszczepu mózgu. IBM, na przykład, zdołał rozwiązać problem słabej identyfikacji płci, który odkrył Joy Buolamwini, budując nowy zestaw treningowy z większą liczbą zdjęć czarnoskórych kobiet. Google rozwiązało swoje wyzwanie z gorylami w odwrotny sposób: usuwając zdjęcia goryli z zestawu treningowego. Żadne rozwiązanie nie jest ogólne; oba są zamiast tego hackami, zaprojektowanymi, aby oślepić analizę danych, aby postępować właściwie, bez naprawiania podstawowych problemów. Podobnie można rozwiązać problemy Tesli z wjeżdżaniem na pojazdy uprzywilejowane zatrzymane na autostradach poprzez dodanie lepszych czujników i uzyskanie odpowiednio oznakowanego zestawu przykładów, ale kto ma powiedzieć, że będzie to działać z lawetami, które zdarzyły się zatrzymać na poboczu Autostrada? Albo pojazdy budowlane? Google może naprawić problem polegający na tym, że obrazy „matki" są prawie całe białe, ale potem problem pojawia się ponownie w przypadku „babci". Ale tak długo, jak dominujące podejście koncentruje się na wąskiej sztucznej inteligencji i coraz większych zestawach danych, pole może utknąć w

nieskończoność, wyszukując krótkoterminowe łatki danych dla określonych problemów, nigdy nie zajmując się podstawowymi wadami, które powodują te problemy są tak powszechne. Potrzebujemy systemów wystarczająco inteligentnych, aby przede wszystkim uniknąć tych błędów. Dzisiaj wydaje się, że prawie wszyscy pokładają swoje nadzieje w głębokim uczeniu się. Co, jak wyjaśnimy za chwilę, uważamy, że to błąd.