Uwzględnienie etycznych aspektów nauki o danych

Według ankiety Gartnera CIO Agenda Survey z 2018 r., 85% projektów AI do 2020 r. przyniesie błędne wyniki z powodu stronniczości danych, algorytmów lub zespołów programistycznych. To poważna postać, którą należy się zająć. Pomyśl o tym: jeśli coraz więcej firm i organizacji staje się opartych na danych i sztucznej inteligencji, a także automatyzuje swoje działania w oparciu o technologie sztucznej inteligencji, którym nie można ufać, oznacza to, że na horyzoncie pojawiają się kłopoty - nie tylko z ewolucją sztucznej inteligencji , ale dla całego społeczeństwa. W tym kontekście zajęcie się etycznymi aspektami sztucznej inteligencji ma fundamentalne znaczenie i będzie moim celem w tym rozdziale. Ale chcę też wyjaśnić, że nie powinieneś zaczynać myśleć o etyce dopiero wtedy, gdy zaczniesz wdrażać swoją strategię analizy danych. Perspektywa etyczna jest w rzeczywistości niezwykle ważna do rozważenia od samego początku – to znaczy od momentu rozpoczęcia projektowania modeli biznesowych, architektury, infrastruktury i sposobów pracy oraz budowania samych zespołów. W tym rozdziale wyjaśniono podstawy, które należy wziąć pod uwagę zarówno z perspektywy strategicznej, jak i praktycznej.

Wyjaśnienie etyki AI

Do czego więc właściwie odnosi się etyka AI i które obszary są ważne, aby wzbudzić zaufanie do danych i algorytmów? Cóż, ta koncepcja ma wiele aspektów, ale istnieje pięć podstaw, na których można polegać;

* Bezstronne dane, zespoły i algorytmy. Odnosi się to do znaczenia zarządzania nieodłącznymi uprzedzeniami, które mogą wynikać ze składu zespołu programistów, jeśli nie ma dobrej reprezentacji płci, rasy i płci. Dane i metody szkoleniowe muszą być jasno określone i uwzględnione w projekcie AI. Zdobywanie spostrzeżeń i potencjalne podejmowanie decyzji w oparciu o model, który jest w jakiś sposób stronniczy (na przykład tendencja do nierówności płci lub postaw rasistowskich), nie jest czymś, co chcesz osiągnąć.

* Wydajność algorytmu. Wyniki decyzji AI powinny być zgodne z oczekiwaniami interesariuszy, że algorytm działa na pożądanym poziomie precyzji i spójności oraz nie odbiega od celu modelu. Gdy modele są następnie wdrażane w środowisku docelowym w sposób dynamiczny i nadal trenują i optymalizują wydajność modelu, model dostosuje się do potencjalnych nowych wzorców danych i preferencji i może zacząć odbiegać od pierwotnego celu. Dlatego niezbędne jest ustalenie wystarczających polityk, aby szkolenie modelowe było zgodne z celami.

* Odporna infrastruktura. Upewnij się, że dane wykorzystywane przez komponenty systemu AI oraz sam algorytm są zabezpieczone przed nieautoryzowanym dostępem, uszkodzeniem i/lub atakiem adwersarza.

* Przejrzystość użytkowania i zgoda użytkownika. Użytkownik musi zostać wyraźnie powiadomiony o interakcji z AI i musi mieć możliwość wybrania poziomu interakcji lub całkowitego odrzucenia tej interakcji. Odnosi się również do znaczenia uzyskania zgody użytkownika na gromadzone i wykorzystywane dane. Wprowadzenie ogólnego rozporządzenia o ochronie danych (RODO) w UE wywołało dyskusje w USA wzywające do podobnych środków, co oznacza, że świadomość wagi danych osobowych, a także potrzeba ochrony tych informacji powoli się poprawia. Tak więc, nawet jeśli dane są gromadzone w sposób bezstronny, a modele są budowane w sposób bezstronny, nadal możesz spotkać się z etycznie trudnymi sytuacjami (lub nawet złamać prawo), jeśli używasz danych osobowych bez odpowiednich uprawnień.

* Modele do wyjaśnienia. Odnosi się to do potrzeby, aby metody szkoleniowe i kryteria decyzyjne AI były łatwe do zrozumienia, udokumentowane i łatwo dostępne do oceny i walidacji przez ludzi. Odnosi się do sytuacji, w których zadbano o to, aby algorytm, będący częścią inteligentnej maszyny, wytwarzał działania, którym ludzie mogą zaufać i które są łatwe do zrozumienia. Przeciwieństwem wyjaśnialności AI jest traktowanie algorytmu jako czarnej skrzynki, w której nawet projektant algorytmu nie jest w stanie wyjaśnić, dlaczego sztuczna inteligencja doszła do określonego spostrzeżenia lub decyzji.

Dodatkowa kwestia etyczna, która ma bardziej techniczny charakter, dotyczy odtwarzalności wyników poza środowiskiem laboratoryjnym. Sztuczna inteligencja jest wciąż niedojrzała, a większość prac badawczo-rozwojowych ma charakter eksploracyjny. Nadal istnieje niewielka standaryzacja dotycząca uczenia maszynowego/sztucznej inteligencji. Pojawiają się de facto zasady rozwoju AI, ale powoli i nadal są one w dużej mierze napędzane przez społeczność. Dlatego musisz upewnić się, że wszelkie wyniki algorytmu są rzeczywiście odtwarzalne – co oznacza, że uzyskasz takie same wyniki w rzeczywistym, docelowym środowisku, jak nie tylko w środowisku laboratoryjnym, ale także między różnymi środowiskami docelowymi (pomiędzy różnymi operatorami w sektorze telekomunikacyjnym , na przykład.)

Adresowanie godnej zaufania sztucznej inteligencji

Jeśli dane, do których potrzebujesz dostępu, aby zrealizować swoje cele biznesowe, można uznać za niepoprawne etycznie, jak sobie z tym radzisz? Łatwo powiedzieć, że aplikacje nie powinny zbierać danych o rasie, płci, niepełnosprawności lub innych chronionych klasach. Ale faktem jest, że jeśli nie zbierzesz tego typu danych, będziesz miał problem ze sprawdzeniem, czy Twoje aplikacje są rzeczywiście uczciwe wobec mniejszości. Algorytmy uczenia maszynowego, które uczą się na podstawie danych, staną się tak dobre, jak dane, na których działają. Niestety, wiele algorytmów okazało się całkiem dobrych w ustalaniu własnych odpowiedników dla rasy i innych klas, w sposób sprzeczny z tym, co wielu uważa za właściwe ludzkie myślenie etyczne. Twoja aplikacja nie byłaby pierwszym systemem, który mógłby okazać się niesprawiedliwy, pomimo najlepszych intencji jego twórców. Ale żeby było jasne, ostatecznie Twoja firma będzie odpowiedzialna za działanie swoich algorytmów, a (miejmy nadzieję) przepisy dotyczące uprzedzeń w przyszłości będą bardziej rygorystyczne niż obecnie. Jeśli firma nie przestrzega praw i przepisów lub granic etycznych, koszty finansowe mogą być znaczne – a być może nawet gorzej, ludzie mogą całkowicie stracić zaufanie do firmy. Może to mieć poważne konsekwencje, od klientów porzucających markę, przez pracowników tracących pracę, po ludzi idących do więzienia. Aby uniknąć tego typu scenariuszy, należy wcielić w życie zasady etyczne, a aby tak się stało, pracownicy muszą mieć możliwość i zachęcać pracowników do etycznego postępowania w codziennej pracy. Powinni umieć rozmawiać o tym, co tak naprawdę oznacza etyka w kontekście celów biznesowych i jakie koszty dla firmy mogą być w ich imieniu znoszone. Muszą też być w stanie przynajmniej przedyskutować, co by się stało, gdyby rozwiązanie nie mogło zostać wdrożone w etycznie poprawny sposób. Czy taka realizacja wystarczyłaby do jej zakończenia? Ogólnie rzecz biorąc, naukowcy zajmujący się danymi uważają, że ważne jest dzielenie się najlepszymi praktykami i artykułami naukowymi na konferencjach, pisanie postów na blogach oraz opracowywanie technologii i algorytmów open source. Jednak problemy, takie jak uzyskanie świadomej zgody, nie są omawiane tak często. Nie jest tak, że problemy nie są rozpoznawane lub rozumiane; są po prostu postrzegane jako mniej warte dyskusji. Zamiast pozwalać, aby takie nastawienie trwało, firmy powinny aktywnie zachęcać (a nie tylko pozwalać) na więcej dyskusji na temat uczciwości, właściwego wykorzystania danych i szkód, jakie może wyrządzić niewłaściwe wykorzystanie danych. Niedawne skandale związane z naruszeniami bezpieczeństwa komputerowego pokazały konsekwencje chowania głowy w piasek: wiele firm, które nigdy nie poświęciły czasu na wdrożenie dobrych praktyk i zabezpieczeń, teraz płaci za to zaniedbanie szkodą na reputacji i finansach.

Ważne jest, aby dochować takiej samej należytej staranności, jaka jest obecnie stosowana w kwestiach bezpieczeństwa, gdy myślimy o kwestiach takich jak uczciwość, odpowiedzialność i niezamierzone konsekwencje wykorzystania danych. Nigdy nie będzie możliwe przewidzenie wszystkich niezamierzonych konsekwencji takiego użycia i tak, możliwość przewidywania przyszłości jest ograniczona. Ale można było łatwo przewidzieć wiele niezamierzonych konsekwencji. (Funkcja przeglądu roku na Facebooku, która wydawała się robić wszystko, aby przypomnieć użytkownikom Facebooka o śmierci w rodzinie i innych bolesnych wydarzeniach, jest doskonałym przykładem.) Słynne motto Marka Zuckerberga: „Ruszaj się szybko i niszcz rzeczy" jest niedopuszczalne, jeśli nie zostało to przemyślane pod kątem tego, co może się zepsuć. Liderzy firm powinni nalegać, aby mogli rozważać takie aspekty – i zatrzymywać linię produkcyjną, gdy coś pójdzie nie tak. Pomysł ten wywodzi się z metody produkcyjnej Andona Toyoty: każdy pracownik linii montażowej może zatrzymać linię, jeśli zobaczy, że coś jest nie tak. Linia nie uruchamia się ponownie, dopóki problem nie zostanie rozwiązany. Pracownicy nie muszą obawiać się konsekwencji ze strony kierownictwa za zatrzymanie linii; cieszą się zaufaniem i oczekuje się od nich odpowiedzialnego zachowania. Co by to znaczyło, gdybyś mógł to zrobić za pomocą funkcji produktu lub algorytmów AI/ML? Gdyby ktoś na Facebooku mógł powiedzieć: „Czekaj, otrzymujemy skargi na przegląd roku" i wycofał go z produkcji, Facebook byłby teraz w znacznie lepszej sytuacji z etycznego punktu widzenia. Oczywiście to duża, skomplikowana firma, z dużym, skomplikowanym produktem. Ale to samo dotyczy Toyoty i tam zadziałało. Kwestią kryjącą się za wszystkimi tymi obawami jest oczywiście kultura korporacyjna. Środowiska korporacyjne mogą być wrogo nastawione do wszystkiego innego niż krótkoterminowa rentowność. Jednak w czasach, gdy publiczna nieufność i rozczarowanie są na najwyższym poziomie, etyka zamienia się w dobrą inwestycję korporacyjną. Kierownictwo wyższego szczebla dopiero zaczyna to dostrzegać, a zmiany w kulturze korporacyjnej nie nastąpią szybko, ale jasne jest, że użytkownicy chcą współpracować z firmami, które traktują ich i ich dane w sposób odpowiedzialny, a nie tylko jako potencjalny zysk lub zaangażowanie, zmaksymalizowany. Firmy, które odniosą sukces w zakresie etyki AI, to te, które tworzą przestrzeń dla etyki w swoich organizacjach. Oznacza to umożliwienie naukowcom zajmującym się danymi, inżynierom danych, programistom i innym specjalistom ds. danych „zajęcie się etyką" w praktyce. Nie chodzi o zatrudnianie wyszkolonych etyków i przydzielanie ich do swoich zespołów; chodzi o to, by każdego dnia żyć wartościami etycznymi, a nie tylko o nich mówić. To właśnie oznacza „robić dobrą analizę danych".

Przedstawiamy etykę według projektu

Jaki jest najlepszy sposób podejścia do wdrażania etyki AI już w fazie projektowania? Czy może być dostępna lista kontrolna? Teraz, kiedy o tym wspomniałeś, jest jeden i znajdziesz go w Wielkiej Brytanii. Tamtejszy rząd uruchomił ramy etyki danych, zawierające podręcznik etyki danych. W ramach inicjatywy wyodrębnili siedem odrębnych zasad dotyczących etyki AI. Zeszyt ćwiczeń, który wymyślili, składa się z szeregu pytań otwartych, które mają na celu zbadanie, czy przestrzegasz tych zasad. Trzeba przyznać, że jest wiele pytań – a dokładnie 46, co jest zbyt dużą liczbą dla analityka danych, aby stale śledzić i skutecznie włączać do codziennej rutyny. Aby takie pytania były naprawdę przydatne, muszą być osadzone nie tylko w rozwojowych sposobach pracy, ale także jako część infrastruktury i wsparcia systemów data science. Nie chodzi tylko o umożliwienie praktycznego przestrzegania zasad etycznych w codziennej pracy i udowodnienie, że firma działa zgodnie z zasadami etyki – firma musi również stać za tymi ambicjami i uwzględniać je jako część swojego kodeksu postępowania . Jednak gdy firma mówi o dodaniu etyki AI do swojego kodeksu postępowania, wartość nie wynika z samej obietnicy, ale raczej wynika z procesu, jaki ludzie przechodzą podczas jej opracowywania. Ludzie, którzy pracują z danymi, zaczynają teraz prowadzić dyskusje na szeroką skalę, które nigdy nie miałyby miejsca jeszcze dziesięć lat temu. Ale same dyskusje nie zakończą ciężkiej pracy. Istotne jest, aby nie tylko mówić o tym, jak korzystać z danych w sposób etyczny, ale także o etycznym korzystaniu z danych. Zasady muszą zostać

wprowadzone w życie! Oto krótsza lista pytań, które należy rozważyć, gdy Ty i Twoje zespoły ds. analityki danych współpracujecie, aby uzyskać wspólne i ogólne zrozumienie tego, co jest potrzebne do rozwiązania problemów etycznych związanych z AI:

* Hakowanie: W jakim stopniu zamierzona technologia AI jest podatna na hakowanie, a tym samym potencjalnie podatna na nadużycia?

* Dane treningowe: Czy przetestowałeś swoje dane treningowe, aby upewnić się, że są uczciwe i reprezentatywne?

* Stronniczość: Czy Twoje dane zawierają możliwe źródła uprzedzeń?

* Skład zespołu: Czy skład zespołu odzwierciedla różnorodność opinii i środowisk?

* Zgoda: Czy potrzebujesz zgody użytkownika na zbieranie i wykorzystywanie danych? Czy masz mechanizm zbierania zgód od użytkowników? Czy jasno wyjaśniłeś, na co użytkownicy wyrażają zgodę?

* Odszkodowanie: Czy oferujesz zwrot kosztów, jeśli ludzie ucierpią z powodu wyników Twojej technologii AI?

* Hamulec awaryjny: czy możesz wyłączyć to oprogramowanie w środowisku produkcyjnym, jeśli zachowuje się źle?

* Przejrzystość i uczciwość: czy wykorzystywane dane i algorytmy sztucznej inteligencji są zgodne z wartościami korporacyjnymi dotyczącymi technologii, takimi jak zachowanie moralne, szacunek, uczciwość i przejrzystość? Czy przetestowałeś uczciwość w odniesieniu do różnych grup użytkowników?

* Wskaźniki błędów: czy przetestowałeś różne poziomy błędów wśród różnych grup użytkowników?

* Wydajność modelu: Czy monitorujesz wydajność modelu, aby zapewnić, że oprogramowanie pozostanie sprawiedliwe w czasie? Czy można mu zaufać, że będzie działał zgodnie z zamierzeniami, nie tylko podczas wstępnego szkolenia lub modelowania, ale także podczas jego ciągłego „uczenia się" i ewolucji?

* Bezpieczeństwo: Czy masz plan ochrony i zabezpieczenia danych użytkownika?

* Odpowiedzialność: Czy istnieje wyraźna linia odpowiedzialności wobec jednostki i jasność dotycząca sposobu działania sztucznej inteligencji, używanych przez nią danych i stosowanych ram decyzyjnych?

* Projekt: Czy projekt AI uwzględniał lokalny i makro-społeczny wpływ, w tym jego wpływ na finansowe, fizyczne i psychiczne samopoczucie ludzi i naszego środowiska naturalnego?