13 Bezpieczeństwo SI z perspektywy pierwszej osoby

Edward Frenkel

SPIS TREŚCI

Chociaż nie jestem ekspertem w dziedzinie bezpieczeństwa sztucznej inteligencji (SI), nie jestem też w tej kwestii zupełnym laikiem. Jako matematyk i profesor na Berkeley University of California, przez lata byłem związany z wieloma różnymi zagadnieniami informatyki, które dotyczyły SI. Pomimo że moje obecne badania dotyczą programu Langlands, interfejsu matematyki i fizyki kwantowej [1, 2], kiedy byłem studentem, zapisałem się na kurs matematyki stosowanej, który obejmował znaczny zakres informatyki i matematyki stosowanej. Moje pierwsze projekty badawcze dotyczyły w rzeczywistości opracowania algorytmów decyzyjnych w diagnostyce medycznej [2, rozdz. 12] i nadal interesuję się takimi tematami, jak analiza danych i uczenie maszynowe. Interesuje mnie również sposób, w jaki technologia zmienia nasze społeczeństwo i co oznacza to dla ludzkości. Bezpieczeństwo SI z pewnością stanowi jedno z najważniejszych i najtrudniejszych wyzwań w tej dziedzinie. Kiedy zostałem zaproszony do napisania tego rozdziału, przyjąłem zaproszenie właśnie z powodu tych zainteresowań, a także dlatego, że chciałem wspomnieć o jednym istotnym aspekcie tego wyzwania, który rzadko, jeśli w ogóle, jest omawiany.

Nazywam to „perspektywą pierwszej osoby", w przeciwieństwie do i jako uzupełnienie „perspektywy trzeciej osoby", która jest dziś znacznie bardziej rozpowszechniona w badaniach naukowych. Perspektywa trzeciej osoby oznacza, że analizuje się obiektywne zjawiska postrzegane „na świecie" poza samym sobą, z punktu widzenia niezależnego obserwatora, właśnie „osoby trzeciej". Perspektywa pierwszej osoby sugeruje, że obserwujemy także siebie, a zwłaszcza relacje ze światem i z badanym zjawiskiem. Ta perspektywa musi zatem zawierać elementy subiektywne.

Przez wieki naukowcy postrzegali świat jako zbiór obiektów oddziałujących ze sobą, ale niezależnych od obserwatora, i dlatego wszelkie wskazówki subiektywne były krytykowane i odrzucane. Jednak od początku XX wieku, z takimi przełomowymi osiągnięciami, jak mechanika kwantowa, ogólna teoria względności i twierdzenia o niekompletności Gödla, nauka uczy nas, że tak naprawdę nie możemy oddzielić obserwatora od obserwowanego obiektu. W ten sposób zniesiono tabu perspektywy pierwszej osoby i stopniowo coraz większa liczba naukowców zaczęła na to zwracać uwagę. Jednak naukowcy wciąż mają przed sobą długą drogę, by zaakceptować i włączyć perspektywę subiektywną i pierwszoosobową do naszego

światopoglądu opartego na wiedzy. Co więcej, uważam, że musimy to zrobić precyzyjnie już dzisiaj ze względu na wyzwania związane z bezpieczeństwem dotyczącym rozwoju SI.

Przypomina to słowa [3] wielkiego amerykańskiego filozofa, psychologa i lekarza Williama Jamesa, którego dzieła wpłynęły między innymi na Bertranda Russella i Nielsa Bohra [4]:

> Nic w duchu i zasadach nauki nie przeszkadza w nauce radzenia sobie ze światem, w którym siły osobowe są punktem wyjścia nowych efektów. Jedyną formą rzeczy, z którą się bezpośrednio spotykamy, jedynym doświadczeniem, które konkretnie mamy, jest nasze życie osobiste. Jedyną kompletną kategorią naszego myślenia jest kategoria osobowości, a każda inna kategoria jest jednym z jej abstrakcyjnych elementów. I to systematyczne zaprzeczanie naukowej części osobowości jako warunku wydarzeń, to rygorystyczne przekonanie, że ze względu na swoją zasadniczą i najgłębszą naturę nasz świat jest światem ściśle bezosobowym, może, jak się wydaje, wraz z upływem wiru czasu, okazać się na tyle mylne, że nasi potomkowie będą wielce zaskoczeni w naszej chwalonej nauce, zaniedbaniem, które ich zdaniem najczęściej sprawia, że wygląda ona na pozbawioną perspektywy i krótką.

Jak to wpływa na bezpieczeństwo związane z rozwojem SI? Odpowiedź jest taka, że ostatecznie to ludzie programują maszyny. Nasze wcześniejsze doświadczenia, nasze obawy, niepewność i inne „problemy" i „zawieszenia" informują i wpływają na nasze przekonania i nasze zachowanie, a zatem mają znaczenie dla tego, co robią badacze i praktycy SI, kiedy tworzą i utrzymują systemy i protokoły SI. Dlatego też, jeśli badacz lub programista SI nie jest w pełni świadomy tego, czym kieruje się SI, może nie być w stanie wykonywać swoich obowiązków w pełni, tak aby zapewnić bezpieczeństwo kontrolowanego systemu i ludzi, którzy są z nim związani.

Można to zilustrować następującą analogią odnoszącą się do tragicznej historii lotu 9525 Germanwings, który rozbił się w Alpach 25 marca 2015 roku, w wyniku czego zginęli wszyscy pasażerowie i załoga. Dowody wskazywały, że wkrótce po starcie drugi pilot zablokował pilotowi dostęp do kokpitu i celowo rozbił samolot o górę. Co więcej, ujawniono później, że drugi pilot był leczony z powodu skłonności samobójczych, ale ukrył tę informację przed linią lotniczą. W ten sposób mógł ominąć przepisy bezpieczeństwa linii lotniczych i popełnić samobójstwo, rozbijając samolot i zabierając ze sobą 150 niewinnych ludzi. Nie trzeba dodawać, że wszyscy ci ludzie przed wejściem na pokład założyli, że ich bezpieczeństwo będzie zapewnione zgodnie ze zwykłymi protokołami bezpieczeństwa. Jednak protokoły te były bezsilne wobec samotnej osoby z samobójczymi zamiarami.

Kuszące jest odrzucenie podobnego scenariusza w przypadku bezpieczeństwa związanego z SI jako zbyt daleko idącego. Zastanówmy się jednak nad następującą odmianą. Załóżmy, że osoba, której zadaniem jest zapewnienie bezpieczeństwa systemu SI, potajemnie wierzy, że ludzie są strasznie i nieodwracalnie wadliwi jako gatunek, a ich jedyną pozytywną rolą w ewolucji jest stworzenie i umocnienie nadrzędnej „rasy robotów", która następnie przypuszczalnie ujarzmi lub eksterminuje „bezużytecznych ludzi". Pomysły tego rodzaju są obecnie poważnie rozpatrywane i dyskutowane przez wielu znanych i szanowanych badaczy SI (patrz np. [5]), a wiele osób wydaje się bardzo przychylnych tym przemyśleniom.

Pomyślmy o tym. Odkładając na bok prawdziwość tych pomysłów, jasne jest, że jeśli osoba mająca takie myśli jest zaangażowana w bezpieczeństwo SI, to może ona być skłonna do wykorzystania swojej pozycji w celu sabotażu nawet najbardziej rygorystycznych protokołów bezpieczeństwa SI, tym samym stwarzając zagrożenie dla innych ludzi. Co więcej, osoba taka nie miałaby żadnych moralnych zastrzeżeń do tego, ponieważ wierzyłaby, że spełnia to, co postrzega jako swój „obowiązek ewolucyjny".