Sztuczną inteligencję da się oszukać

14 grudnia 2020

Naklejka na znaku drogowym może wprowadzić w błąd autonomiczny samochód, a zmylony zakodowanym sygnałem asystent głosowy może zlecić przelew naszych pieniędzy na wybrane konto. Ataki na systemy, które potrafią same się uczyć, mają więcej wspólnego z codziennym życiem, niż mogłoby się wydawać.

SYSTEM OSZUKUJE SYSTEM

„Sztuczna inteligencja nas obserwuje, słyszy i uczy się nas – czy tego chcemy, czy nie. W świecie Web 3.0 inteligentne systemy są wykorzystywane na każdym portalu internetowym. W zależności od tego, czego szukamy w sieci, co kupujemy, o czym mówimy i piszemy przez telefon, są nam sugerowane wybrane dla nas treści. Portale społecznościowe proponują nam znajomych, serwisy sprzedażowe podsuwają produkty i usługi, jakich „potrzebujemy"" – mówi dr inż. Rafał Kasprzyk z Wojskowej Akademii Technicznej w Warszawie.

Zaznacza jednak, że te dobrze działające urządzenia można wprowadzić w błąd. Sposób na oszukanie sztucznej inteligencji określany jest w języku naukowym jako antagonistyczne, czyli przeciwstawne uczenie się maszyn. W skrócie – drugi system uczy się, jak działa pierwszy, który również sam się uczy.

Dzięki temu procederowi może się zdarzyć, że maszyna „połknie" na wejściu złośliwie spreparowane dane. Dla człowieka takie dane wyglądają na poprawne i nie wprowadzają go w błąd. Maszyna jednak daje się nabrać. Naukowiec wyjaśnia to na przykładzie autonomicznych samochodów.

Jak opisuje, auta takie jak Tesla bardzo dobrze trzymają się pasów i świetnie rozpoznają znaki drogowe – nawet te na

naszych drogach – bardzo liczne. Choćby warunki atmosferyczne były złe, a znak stop byłby pomalowany przez kogoś na zielono, sensory uruchomią mechanizmy ochrony przed wypadkiem.

Jeśli jednak system zostanie w sposób przemyślany zaatakowany, wtedy – na przykład – wystarczy nakleić niepozorny element na znaku STOP, aby samochód odebrał ten znak jako informację: „droga z pierwszeństwem przejazdu”. Nietrudno wyobrazić sobie konsekwencje w postaci niewyjaśnionej serii kraks.

DWIE DROGI UCZENIA (SIĘ) MASZYN

Dr inż. Kasprzyk tłumaczy, że przy budowie inteligentnych maszyn można wykorzystać wiedzę eksperta. Człowiek – specjalista z danej dziedziny – wskazuje, jak powinna funkcjonować maszyna. Następnie informatyk „tłumaczy” to na kod w wybranym języku programowania maszyn. Opracowuje jasne reguły, które wyznaczają sposób działania maszyny.

Można to porównać do sytuacji, w której sami programujemy ekspres do kawy, żeby – bez ustawiania wszystkiego za każdym razem – serwował nam wymarzoną kawę o ulubionej porze.

Drugi sposób jest bardziej abstrakcyjny, bo mamy w nim do czynienia z ogromną ilością danych. Żaden ekspert ani grono specjalistów, nie jest w stanie ogarnąć umysłem mechanizmów zarządzania nimi. A jednak – choć reguły, jakimi rządzi się maszyna nie są jawne i znane naukowcom – system, zaprogramowany przez człowieka na samodzielne uczenie się, działa prawidłowo, niejednokrotnie znacznie lepiej niż sam człowiek.

To podejście można z kolei porównać do sytuacji, w której ekspres do kawy automatycznie przygotowuje właściwą kawę. Bierze pod uwagę godzinę, temperaturę i wilgotność w pomieszczeniu, dzień tygodnia i – może jeszcze coś więcej? A uczy się w oparciu o to, w jaki sposób użytkownik do tej pory wykorzystywał go w różnych porach dnia, miesiąca, roku.

Istnieje wiele algorytmów uczenia maszynowego, czyli metod, według których ludzie budują maszyny „karmiące się" niezliczoną ilością danych. Dalej „pałeczkę" przejmuje system, za którym ludzki umysł już nie może nadążyć. Informacje zbierane są przez sensory otaczające nas w świecie rzeczywistym i wirtualnym. Inteligentne maszyny przetwarzają te dane i szukają między nimi zależności. Na ich podstawie dalej się uczą.

Naukowcy potrafią nauczyć inteligentne maszyny, jak identyfikować osoby podejrzane o terroryzm. W projekcie Pentagonu powstał system „Znawca" (Maven), który potrafi namierzać, rozpoznawać, a następnie śledzić obiekty na obrazach rejestrowanych przez satelity lub bezzałogowe statki powietrzne.

Naukowcy – m.in. na Wydziale Cybernetyki WAT szukają możliwych sposobów na wprowadzanie maszyn w błąd. Znając je – będą mogli temu zapobiec.

Autorstwo: Karolina Duszczyk
Źródło: NaukawPolsce.PAP.pl