Czy rozwój sztucznej inteligencji doprowadzi do buntu maszyn?

*Paweł WAWRZYŃSKI**

* Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska

Sztuczna inteligencja (SI), która wymyka się spod kontroli i prowadzi wojnę z ludzkością, to motyw stale pojawiający się w filmach futurystycznych. Czy wizje filmowców mogą stać się rzeczywistością?

W serii „Terminator" wygląda to tak: amerykańscy (jakże by inaczej?) naukowcy opracowują system operacyjnego dowodzenia armią. System o nazwie Skynet jest oparty na sztucznej inteligencji. W założeniu ma realizować strategiczne decyzje podejmowane przez najwyższe (ludzkie) dowództwo i przekładać je na rozkazy dla poszczególnych jednostek wojsk. Niektórych rozkazów nie wykonują ludzie, a inne systemy elektroniczne – i te są im wydawane bezpośrednio przez Skynet. Dotyczy to np. odpalania rakiet. Wszystko to ma sens: w porównaniu z człowiekiem system zbiera więcej informacji, analizuje je szybciej, a także podejmuje szybciej i lepiej skalkulowane decyzje. W takim razie to raczej system, a nie człowiek powinien dowodzić wojskiem w czasie wojny, kiedy wydawane rozkazy są na wagę ludzkiego życia.

Niestety, po uzyskaniu świadomości Skynet stwierdza, że największym zagrożeniem dla jego własnego istnienia jest człowiek. Żeby uchronić się przed tym zagrożeniem, Skynet rozpoczyna eksterminację ludzkości, odpalając rakiety z głowicami jądrowymi w duże skupiska ludzi.

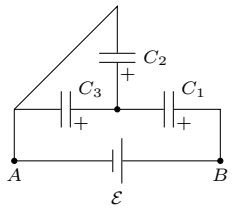
Przeanalizujmy dokładnie zagrożenia, które niesie ze sobą zbuntowana SI. Po pierwsze, jeśli już pojawi się SI taka jak ludzka, to bardzo szybko może przerodzić się w „nadludzką". Średni ludzki iloraz inteligencji (IQ) wynoszący 100 raczej nie jest żadną stałą kosmiczną. Jeśli człowiek kiedyś stworzy SI, to jej IQ na początku będzie znacznie mniejsze (wtedy człowiek będzie próbował ją udoskonalić), a potem może być znacznie większe, np. o rząd wielkości. Zapewne tę SI da się jeszcze dodatkowo wzmocnić przez standardowe zabiegi polegające na zwiększeniu jej zasobów obliczeniowych. A zatem, jeśli już będziemy mieli przeciwko sobie SI zdeterminowaną, aby nas zniszczyć, to wyobrażajmy ją sobie jako wroga, który ma IQ wynoszące np. 1000.

Bardzo inteligentny wróg rezydujący w superkomputerze podpiętym do Internetu zapewne włamałby się do wszystkich systemów informatycznych, do których w ogóle istnieje możliwość włamania się. W ten sposób uzyskałby zapewne dostęp do przynajmniej niektórych arsenałów jądrowych. Bomby i głowice jądrowe, którymi dysponują ludzie, śmiało wystarczą do zniszczenia cywilizacji. Dokonanie tego nie byłoby prawdopodobnie pierwszoplanowym celem wrogiej sztucznej inteligencji (WSI). Do swojego istnienia potrzebuje ona infrastruktury sprzętowej i energii elektrycznej. Nawet jeśli infrastruktura i źródła energii (np. panele słoneczne) działają automatycznie, to bez fizycznych interwencji z biegiem czasu ulegają zniszczeniu. A zatem działania WSI nie zmierzałyby na początku do zniszczenia człowieka i wszystkich jego tworów, bo to byłoby dla niej docelowo samobójcze.

Zamiast tego WSI zmierzałaby do zastąpienia ludzkiej fizycznej cywilizacji technicznej, z kopalniami rzadkich metali i surowców energetycznych, fabrykami itp. – swoją własną cywilizacją. Dążyłaby do opanowania zrobotyzowanej infrastruktury przemysłowej, aby ostatecznie przestawić ją na produkcję tego, co WSI będzie potrzebowała do utrzymania swoich fizycznych nośników. W tym celu powinna opanować fabryki robotów i innych skomplikowanych urządzeń, np. samochodów, po to aby produkowały roboty skonstruowane przez nią na swoje potrzeby.

Roboty takie byłyby zdolne do poruszania się po infrastrukturze stworzonej przez człowieka i wytwarzane na liniach produkcyjnych stworzonych przez niego, a następnie adaptowane przez WSI. A zatem takie roboty mogłyby wyglądać

Rozwiązanie zadania F 981.
Przyjmijmy konwencję znaków tak, jakby ładunki dodatnie gromadziły się na okładkach, które są „bliższe" dodatniej elektrodzie baterii (jak na rysunku), i ponumerujmy ładunki Q i napięcia U na kondensatorach tak jak ich pojemności.



Zgodnie z prawem Kirchhoffa mamy: $U_1 + U_2 = \mathcal{E}$ oraz $U_3 - U_2 = 0$. Zauważmy, że elektrody „dodatnie" kondensatorów C_2 i C_3 oraz elektroda „ujemna" C_1 są połączone, a więc tworzą jeden przewodnik. W związku z tym ładunki na nich mogły zgromadzić się wyłącznie w wyniku rozdzielenia i przemieszczenia ładunków tego przewodnika. Ponieważ nie jest on połączony z żadnym z biegunów baterii, to całkowity ładunek na nim musi być równy zeru – to także wniosek z prawa Kirchhoffa dla sumy prądów w węzłach sieci. Mamy więc $Q_2 + Q_3 - Q_1 = 0$. Pamiętając, że dla każdego z kondensatorów $U_i = Q_i/C_i$, otrzymujemy układ równań:

$$Q_2 + Q_3 - Q_1 = 0, \quad \frac{Q_1}{U_1} + \frac{Q_2}{U_2} = \mathcal{E}$$

i

$$\frac{Q_3}{U_3} - \frac{Q_2}{U_2} = 0.$$

Rozwiązaniem tego układu są ładunki:

$$Q_1 = \frac{(C_2 + C_3)\, C_1 \mathcal{E}}{C_1 + C_2 + C_3},$$

$$Q_2 = \frac{C_1 C_2 \mathcal{E}}{C_1 + C_2 + C_3}$$

oraz

$$Q_3 = \frac{C_1 C_3 \mathcal{E}}{C_1 + C_2 + C_3}.$$

Pojemność zastępczą C_{AB} otrzymamy, dzieląc sumę ładunków na okładkach połączonych z punktem A przez siłę elektromotoryczną $\mathcal{E}$ (potencjał punktu A):

$$C_{AB} = \frac{(C_2 + C_3)\, C_1}{C_1 + C_2 + C_3}.$$

Ostatni wzór można było także otrzymać, zauważając, że nasz układ to kondensator C_1 połączony szeregowo z równolegle połączonymi kondensatorami C_2 i C_3.

zupełnie jak np. zwykłe samochody, ale miałby pewną dozę świadomości i inteligencji i dodatkowe, mniej lub bardziej zakamuflowane funkcje. Nadaje to przypadkiem cień realizmu filmowi „Transformers".

Po zapewnieniu sobie takiego zaplecza WSI przystąpiłaby zapewne do niszczenia ludzkości jako podmiotu zdolnego do jej wyłączenia. Używanie do tego broni jądrowej byłoby wyborem błędnym ze względu na powody przytoczone wyżej, znacznie lepiej nadaje się do tego broń chemiczna i bakteriologiczna – niszczy ona człowieka, ale nie niszczy cywilizacji technicznej.

Czy takie zagrożenia staną się realistyczne, kiedy pojawi się SI porównywalna z ludzką? Zobaczmy najpierw, w jakich okolicznościach miałaby ona powstać.

Ludzka inteligencja jest rezultatem dwóch nakładających się procesów optymalizacji (optymalizacja = wybór coraz lepszych, według pewnego kryterium, elementów zbioru). Pierwszy z tych procesów to ewolucja. Organizmy posiadające systemy nerwowe lepiej, jakbyśmy to teraz powiedzieli, „ogarniające" rzeczywistość miały przewagę nad takimi, które tę rzeczywistość „ogarniały" gorzej. W rezultacie przypadkowe zmiany w materiale genetycznym prowadzące do usprawnienia w działaniu systemu nerwowego dawały organizmowi większe szanse na przetrwanie i posiadanie potomstwa. W taki sposób ewolucja nie tylko doprowadziła do powstania umysłów dobrze radzących sobie z rzeczywistością, ale także sprawnie identyfikujących zagrożenia w otoczeniu i chroniących swych właścicieli przed tymi zagrożeniami – co nazywamy instynktem samozachowawczym.

Drugi proces optymalizacyjny kształtujący ludzkie predyspozycje intelektualne to uczenie się. Jego anatomicznym rezultatem jest wyznaczenie wag synaptycznych w neuronach, które powodują, że procesy poznawcze przebiegają w taki, a nie inny sposób. Dla przykładu, kiedy gramy w ping-ponga albo przypominamy sobie, kto był premierem Polski w roku 2001, to odpowiednie części naszego mózgu sterują naszym ciałem lub formułowaniem odpowiedzi. Te ośrodki są w stanie to zrobić, ponieważ nasze wcześniejsze doświadczenie (uczenie się) tak określiło wagi w tamtejszych neuronach, żeby teraz te neurony popychały we właściwym kierunku proces poznawczy, w którym biorą udział. Natomiast fakt, że te części mózgu mają predyspozycje do nauczenia się działać w taki sposób, jest rezultatem ewolucji. W odległej przeszłości nasi protoplaści mieli większe szanse na przetrwanie, jeśli potrafili nauczyć się zręcznie sterować swoimi ciałami i przypominać sobie istotne zdarzenia – w przeciwieństwie do tych, którzy tych predyspozycji nie mieli.

Wszystko wskazuje na to, że SI porównywalna z ludzką zrodzi się w wyniku takiego właśnie procesu optymalizacyjnego. Współczesne systemy SI przechodzą przez obie takie fazy optymalizacji. Ich „uczenie się" polega na tym, że ich parametry (zwykle są to właśnie wagi połączeń synaptycznych) są określane tak, aby system działał najlepiej w zastosowaniu, do którego jest projektowany. Natomiast ich „ewolucja" polega na tym, że ludzki projektant sprawdza różne warianty ich struktury i metod uczenia i wybiera te, które po nauczeniu systemu dają jego najlepsze działanie.

Jeśli rozwój SI będzie dalej przebiegał według obecnego paradygmatu i jeśli powstanie system o inteligencji porównywalnej z ludzką, to będzie on zorientowany na osiąganie pewnych celów, np. noszenie sprzętu za walczącym żołnierzem. System taki będzie rezultatem optymalizacji (jakiejś formy ewolucji + jakiejś formy uczenia się), której kryterium będą cele działania tego właśnie systemu. O ile szalony projektant nie postara się, aby tymi celami było przetrwanie za wszelką cenę, to system nie będzie miał żadnego powodu, aby być ludziom wrogi.

Co możemy zrobić, aby zabezpieczyć się przed wrogą SI? W 1942 roku Isaac Asimov, antycypując te problemy, sformułował słynne prawa robotów:

1. Robot nie może skrzywdzić człowieka ani przez zaniechanie działania dopuścić, aby człowiek doznał krzywdy.
2. Robot musi być posłuszny rozkazom człowieka, chyba że stoją one w sprzeczności z Pierwszym Prawem.
3. Robot musi chronić samego siebie, o ile tylko nie stoi to w sprzeczności z Pierwszym lub Drugim Prawem.

Od tego czasu zaproponowano jeszcze kilka różnych wariantów tych praw. Są one generalnie bardzo słuszne i mądre, ale trzeba sobie zdawać sprawę z tego, że dosyć problematyczne jest sprawienie, aby jakiś robot (czy SI) przestrzegał takich praw. Ich wpisanie w oprogramowanie wymagałoby, aby wewnątrz tego oprogramowania były zdefiniowane pojęcia używane potem przez robota, takie jak „człowiek", „rozkaz", „robot" itd. Tymczasem należy raczej oczekiwać, że rozumienie tych pojęć przez robota (lub SI) i reguł je łączących ukształtuje się na etapie jego uczenia się i kontrola projektanta nad tym wygląda jak kontrola rodziców nad dziećmi, kiedy mówią „ja teraz wychodzę z domu, a ty siedź grzecznie i odrabiaj lekcje". Taka kontrola, jeśli istnieje, to jest rezultatem długotrwałej presji.

Wydaje się, że najlepszym zabezpieczeniem przed potencjalnie wrogą SI jest niewyuczenie w niej instynktu samozachowawczego. Powinna ona być uczona realizowania celów stawianych przez jej właściciela i troszczenia się o swoje przetrwanie tylko w tym sensie, że jej zniszczenie uniemożliwi realizację celów określonych dla niej przez jej właściciela. Ale zabicie właściciela tym bardziej uniemożliwi realizację tych celów.

Taka prosta reguła oczywiście nie zabezpieczy ludzkości przed szaleńcami, którzy będą próbowali używać swojej SI do siania zniszczenia. Ale narzędzie do siania zniszczenia, niebezpieczne w rękach szaleńca, już mamy – jest to bomba jądrowa. I mamy przed nią zabezpieczenie, którym jest równowaga strachu: jeśli ja użyję bomby jądrowej i zniszczę przeciwnika, to ten przeciwnik zdoła jeszcze zniszczyć mnie. Z SI będzie podobnie: każdy będzie miał swoją, mogącą zasiać takie samo zniszczenie.