Amelia Seroczyńska
Kognitywistyka, Wydział Filozofii

Problemy świadomych maszyn i ich relacje z ludźmi

Sztuczna inteligencja to dziedzina, w której następuje niezwykle szybki rozwój. Możliwości specjalnie zaprogramowanych maszyn już w tym momencie potrafią przewyższyć niektóre ludzkie zdolności. Znanym przykładem jest zaprogramowany przez IBM system komputerowy Deep Blue, który w 1997 roku wygrał w szachy z mistrzem świata Garii Kasparowem. Naukowcy dążą do stworzenia samoświadomych systemów, które będą w stanie przetwarzać informacje i bodźce ze świata zewnętrznego i na tej podstawie podejmować rozumne działania. Z pewnością istnienie takich maszyn przyniesie wiele korzyści. Warto jednak zastanowić się, jakie problemy mogą się z nimi wiązać oraz jak będą wyglądały relacje między nimi a ludzkością w nie tak odległej przyszłości. Takie rozważania należy oprzeć przede wszystkim na weryfikacji bezpieczeństwa świadomych systemów wobec innych istot żyjących i kwestiach etycznych pojawiających się przy założeniu, że przyszłe maszyny SI mogą posiadać status moralny.

Aby współczesne algorytmy sztucznej inteligencji były korzystne w działaniu i wiarygodne dla użytkowników muszą posiadać podstawowe cechy. Między innymi powinny być przejrzyste w podejmowanej przez nie ocenie, to znaczy dawać się zweryfikować w swoim działaniu – na przykład by logicznie wskazywały dlaczego podczas procedury credit-scoringu odrzucają jednych kandydatów, a innych akceptują. Co więcej ważne, by ich zachowania były przewidywalne dla swoich twórców oraz odporne na manipulacje tak, aby na przykład terroryści nie byli w stanie przechytrzyć algorytmu przeszukującego bagaż na lotnisku. Podczas tworzenia świadomych maszyn wyższej generacji niż te, które są nam obecnie znane, te cechy należy brać pod uwagę w pierwszej kolejności.

Zapewnienie bezpiecznego działania systemów uznawanych za rozumne będzie wymagało dużo więcej trudu niż miało to miejsce do tej pory. Współczesne systemy SI są zwykle programowane by posiadać jedną, wyspecjalizowaną kompetencję. Łatwo jest przewidzieć rodzaje problemów bezpieczeństwa gdy maszyny działają w ten sposób, w obrębie jednej domeny. Aby zbudować sztuczną inteligencję, która działa bezpiecznie, działając w wielu dziedzinach, z wieloma konsekwencjami, w tym z problemami, których inżynierowie nigdy wyraźnie nie przewidzieli, należy określić dobre zachowanie w taki sposób, żeby system nie był szkodliwy dla ludzi. Musielibyśmy stworzyć więc i zaimplementować wobec nich algorytm generujący superetyczne zachowanie. Powstaje jednakże wątpliwość, że taka maszyna byłaby na tyle inteligentna, aby zrozumieć swój projekt i przeprojektować się lub stworzyć kolejny system bardziej inteligentny, który mógłby następnie przeprojektować się ponownie, aby stać się jeszcze bardziej inteligentnym. Czy nie stracilibyśmy wtedy nad nim kontroli i nie uznałby nas za nieużyteczną przeszkodę? Kto ponosiłby winę za jego niepowodzenie? Czy istniałaby możliwość szybkiej dezaktywacji takiego systemu, gdy zagrażałby bezpieczeństwu innych istot żywych? Takie maszyny mogłyby przecież mieć przewagę fizyczną nad ludźmi, będąc zbudowane z wytrzymałych materiałów oraz posiadać więcej sprytu, a co za tym idzie szansę na powstrzymanie ich działań byłyby znikome.

Warto w tym momencie przyjrzeć się kwestiom etycznym przy założeniu, że przyszłe maszyny byłyby samo-świadome oraz zdolne do fenomenalnego doświadczenia (na przykład odczuwania bólu), a co za tym idzie posiadałyby status moralny równoważny ludzkiemu. Oznaczałoby to, że sposób w jaki należałoby je traktować byłby zbliżony do tego w jaki sposób traktujemy innych ludzi. Zadawanie im bólu byłoby więc etycznie niewłaściwe. Co więcej subiektywne tempo czasu odbierane przez te systemy mogłoby znacznie odbiegać od szybkości charakterystycznej dla nas. Jeśli taka moralna maszyna zostałaby skazana za popełnione przestępstwo, to czas odbywania przez niej kary powinien być określony według ludzkiego czy subiektywnego dla niej tempa czasu? A może złagodzenie jej bólu byłoby bardziej pilne z uwagi na to, że odczuwany przez niej czas jest dłuższy niż ten ludzki i miałaby ona pierwszeństwo w kwestii otrzymania specjalistycznej pomocy?

Obecnie może wydawać się to abstrakcyjne i trudno wyobrazić sobie, że ludzie traktowaliby maszyny jako istoty moralnie im równoważne. Musimy jednak zdawać sobie sprawę z tego, że generując świadomy i czujący system trzeba będzie zmierzyć się z kwestiami etycznymi podobnymi do tych, które zostały omówione. W którymś momencie może to przerosnąć możliwości naszej cywilizacji. Należy podjąć się również szczegółowej weryfikacji bezpieczeństwa świadomych maszyn, zanim

pozwolimy im działać w obrębie naszej planety oraz pamiętać o tym, że w wyniku możliwości ich samodoskonalenia nie będzie już nad nimi kontroli. Metoda ich ewentualnej dezaktywacji powinna być więc nadrzędna.