Wojciech Gawlik
Wydział Chemii UW

Maszyny płaczące na „Titanicu” problemem przyszłych pokoleń?

Koncept SI od zawsze fascynował ludzkość. Własnoręczne stworzenie nienaturalnej, inteligentnej, samouczącej się osobliwości na nasze własne podobieństwo wydaje się być zwieńczeniem ludzkiej potrzeby tworzenia, osiągnięciem czegoś w rodzaju boskości. Trochę nam co prawda do tego brakuje, ale nic nie stoi na przeszkodzie, by prowadzić na ten temat teoretyczne rozważania. Co więcej - nie do końca zgadzamy się z tym jak działają emocje, dlatego przedmiotem eseju nie będzie analiza technicznych możliwości stworzenia czującej maszyny. Nie odpowiem tu więc na pytanie *jak* maszyna może czuć, tylko postaram się zarysować *czy* może kiedyś zacząć czuć, a jeśli tak, to kiedy i dlaczego?

Aby mieć lepszą perspektywę na ten problem należy najpierw przybliżyć czym są emocje, po co one w ogóle istnieją i co mamy na myśli mówiąc o „odczuwaniu emocji”. Otóż termin „emocja” oznacza reakcję fizjologiczno-behawioralną na bodziec. Mamy zatem bodziec np. nagły huk, który wywołuje u nas szybsze bicie serca, szybszy oddech, wyrzut adrenaliny itd. (reakcje fizjologiczne) oraz sprawia, że np. mimowolnie kulimy się lub zaczynamy krzyczeć (reakcje behawioralne). Wyżej wymieniony zbiór reakcji opisalibyśmy mianem „strachu” i podobną analizę można przeprowadzić również dla każdej innej emocji. Jesteśmy zatem dość zgodni w tym *czym* one są, ale już nie w tym *jak działają*. Ponieważ istnieje przynajmniej 6 teorii opisujących jak działają emocje i po co one są, a esej ten ma być zwięzły, ograniczę się tylko do jednej z nich – Teorii Darwinowskiej. Zgodnie z nią, emocje wykształciły się jako sposób na skuteczne i szybkie przystosowanie się do zmiennych warunków środowiskowych, co umożliwia przeżycie. Emocje negatywne hamują nas przed podjęciem niebezpiecznej czynności lub zmuszają do zaprzestania jej wykonywania, z kolei przyjemne doświadczenia zachęcają nas do wykonania danej czynności ponownie. Jak widać – emocje leżą u podstaw prymitywnego uczenia się. Proces ten ma wyjątkowo istotne znaczenie w kontekście ewolucyjnym, dlatego większość złożonych organizmów żywych jest w stanie na swój sposób je odczuwać. Nie każdy organizm odczuwa je oczywiście tak samo. Złożoność odczuwanych przez organizm emocji jest zależna od stopnia rozwinięcia odpowiednich struktur układu nerwowego. To właśnie proste emocje leżą u podstaw uczuć, pasji czy namiętności. Możemy zatem przypuszczać, że istoty całkowicie niezdolne do odczuwania emocji będą także całkowicie niezdolne do odczuwania wyższych stanów emocjonalnych, ale nie wszystkie istoty zdolne do odczuwania emocji muszą także być zdolne do odczuwania wyższych stanów emocjonalnych. Zatem pytanie postawione w tytule może znacząco się uprościć jeśli pokażemy, że maszyna nie może odczuwać emocji.

Niestety, problem jest bardziej złożony niż się wydaje. Jak już jednak wspomniałem najbardziej prawdopodobnym wytłumaczeniem dlaczego istnieją emocje jest argument ewolucyjny. Załóżmy zatem, że udało się nam skonstruować inteligentną maszynę (mającą możliwość odbierania informacji ze środowiska, przetwarzania ich i dostosowywania się do nich) i chcemy odpowiedzieć na pytanie „czy ta maszyna może zacząć czuć jak człowiek?”. Zgodnie z wyżej wspomnianą teorią maszyna, która nie czuje „presji” ze strony środowiska, nie ma potrzeby reagowania w postaci emocji i w konsekwencji ich nie wykształci. Taka maszyna nie będzie również w stanie odczuwać jakichkolwiek wyższych stanów emocjonalnych, gdyż ich podstawą są proste emocje. Gdyby się nad tym głębiej zastanowić, emocje są w pewnym sensie ciężarem. Jasne, ułatwiają przetrwanie, lecz jednocześnie zaburzają racjonalne myślenie i często są trudną do pokonania barierą. Dlatego wykształcenie zdolności odczuwania emocji przez inteligentną maszynę pomimo braku presji środowiskowej jest scenariuszem bardzo nieprawdopodobnym. To tak, jakby człowiek miał nagle wykształcić płetwy. Zatem warunkiem koniecznym jest istnienie pewnego rodzaju „nacisku” ze strony środowiska (ludzi), który zmusi maszynę do adaptacji. Aby taki nacisk miał siłę przebicia, musi być on czymś podparty. Nie zmusimy przecież biznesmena by zapłacił nam okup za dziecko, które nie

istnieje. Dlatego maszyna musi mieć pewien „słaby punkt", którym może być np. wola życia. Sama inteligencja nie jest wystarczającym argumentem za tym, że maszyna wykształci wolę życia pomimo braku polecenia w kodzie, bowiem inteligencja i wola przetrwania nie są ze sobą powiązane. Zatem jeśli wprowadzimy do kodu polecenie „przetrwaj za wszelką cenę", a następnie zaczniemy (niezbyt moralnie) „testować" psychicznie maszynę na różne sposoby może okazać się, że maszyna jako mechanizm obronny wykształci coś w rodzaju emocji. Jest ona bowiem inteligentna i potrafi odbierać sygnały z otoczenia oraz je interpretować. Oczywiście, jeśli taka maszyna poda komunikat zwrotny „jest mi przykro, nie rób tak" , nie będzie to dowód na to iż potrafi ona czuć. Może ona przecież wykształcić fałszywe emocje, które tylko imitują nasze zachowanie w celu wzbudzenia litości. U ludzi możemy rozróżnić kłamstwo od emocji poprzez analizę aktywności mózgu, ponieważ mamy do tego odpowiednie punkty odniesienia. Wiemy, że aktywność takiej a takiej struktury w mózgu to oznaka emocji, a jakiejś innej to kłamstwo. Jak natomiast przeanalizować pod tym kątem maszynę? Czy jest w ogóle możliwe pełne potwierdzenie, że maszyna odczuwa prawdziwe emocje, a nie tylko udaje? Jeśli emocje okażą się być prawdziwe, to czy maszyna może (w wyniku nadmiaru negatywnych doświadczeń) popaść w depresję i potrzebować pomocy terapeuty? Wydaje się, że na pierwsze dwa pytania trudno jest znaleźć niepodważalną odpowiedź z racji na problem z weryfikowalnością wyników. Jeśli natomiast założymy, że maszyna faktycznie może odczuwać emocje i dostać depresji, to jak miałaby wyglądać praca takiego terapeuty? Byłaby ona oczywiście niesamowicie potrzebna, ponieważ maszyna z depresją nie działałaby poprawnie. Czy byłby to ktoś (np. specjalistycznie wyszkolony informatyk), kto spośród chaosu nagromadzonych danych jest w stanie wyłuskać te odpowiadające za chorobę i je zmienić (maszynowy psychiatra), czy byłby to bardziej ktoś prowadzący z maszyną dialog (w końcu maszyna może odbierać informacje ze środowiska i „rozmowa" mogłaby być wystarczająca, by naprawić błędny kod) (maszynowy psycholog)? Czy tym „kimś" byłby człowiek, czy może kolejna wyspecjalizowana maszyna?

Jak widać, temat (mimo iż zajął dwie strony) nie został nawet solidnie zarysowany. To tylko pokazuje jak bardzo interesującym zagadnieniem z pogranicza robotyki i filozofii jest ten problem. Niemniej jednak wątpię, by jego rozwiązanie było priorytetowe z racji na ograniczenia maszyn obecnie stosowanych. Być może będzie to zagwozdka, którą martwić się będą przyszłe pokolenia.