Sztuczne emocje i świadomość maszynowa

Wstęp

W ciągu ostatniej dekady w sztucznej inteligencji (AI) zauważalnie wzrosło zainteresowanie sztucznymi emocjami i świadomością maszynową, o czym świadczy szereg specjalistycznych konferencji i warsztatów poświęconych tej tematyce. Zainteresowanie to częściowo opiera się na uznaniu, że emocje i świadomość odgrywają pożyteczną rolę u ludzi i innych zwierząt oraz że zrozumienie tych ról i wdrożenie ich modeli na komputerach może pomóc w uczynieniu sztucznych agentów mądrzejszymi. Ale czy maszyny w ogóle mogą mieć emocje i być świadome, a jeśli tak, to w jaki sposób możemy zabrać się za projektowanie takich maszyn? Celem tego rozdziału jest przedstawienie przeglądu prac AI nad emocjami i świadomością maszynową, z myślą o udzieleniu odpowiedzi na te pytania. Rozpoczynając od krótkiego filozoficznego spojrzenia na emocje i świadomość maszynową, aby nadać ramę pracy, rozdział najpierw koncentruje się na sztucznych emocjach, a następnie przechodzi do świadomości maszynowej – co odzwierciedla fakt, że emocje i świadomość były traktowane niezależnie i przez różne społeczności zajmujące się sztuczną inteligencją. Część kończy się omówieniem filozoficznych implikacji badań nad sztuczną inteligencją dla emocji i świadomości.

Perspektywa filozoficzna

Prima facie wydaje się, że badania nad emocjami i świadomością w AI należałoby zacząć od założenia, że faktycznie możliwa jest implementacja emocji i świadomości w artefaktach obliczeniowych. Po co innego zawracać sobie głowę próbą osiągnięcia tego celu, skoro w zasadzie nie można go osiągnąć? Okazuje się, że badacze sztucznej inteligencji zazwyczaj nie byli pod wrażeniem filozoficznych argumentów na temat możliwości lub niemożności stworzenia maszyn replikujących ludzkie stany psychiczne. Raczej zawsze stosowali teoretycznie nieskrępowane podejście do badania możliwych algorytmów i mechanizmów umożliwiających osiągnięcie inteligentnego zachowania. W sztucznej inteligencji istnieją zasadniczo dwa główne podejścia do pytania, czy maszyny rzeczywiście mogą odczuwać emocje (np. jak emocje ludzkie) lub być świadome (np. jak normalny dorosły człowiek w stanie czuwania). Pierwsza to pragmatyczne podejście, które leży u podstaw większości badań nad sztuczną inteligencją i łączy się z pokrewnymi postawami w psychologii: terminy związane z emocjami i „świadomość" są używane w pragmatyczny, operacyjny sposób, który pozwala badaczom czynić postępy bez rozwiązywania wszystkich problemów koncepcyjnych, które nękają te koncepcje . Badacze zajmujący się sztuczną inteligencją, którzy przyjmują takie podejście, przyjrzą się wynikom psychologii pod kątem typów procesów, które psychologowie podejmują, aby leżeć u podstaw ludzkiej aktywności umysłowej lub być w nią zaangażowani, i spróbują sformalizować ich aspekty w sposób algorytmiczny. Celem nie jest tutaj odtworzenie ani modelowanie ludzkiej mentalności w biologicznie lub psychologicznie wiarygodny sposób, ale raczej wykorzystanie wszelkich zasad, które można zaczerpnąć z procesów emocjonalnych lub teorii świadomości, aby poprawić działanie sztucznych czynników (i prawdopodobnie przewyższyć możliwości ludzkie). Druga postawa polega na próbie udoskonalenia, zrewidowania lub zastąpienia koncepcji emocji lub koncepcji świadomości w wyniku próby formalnego określenia procesów, które mogą wdrożyć emocje lub doprowadzić do świadomości. Postawa ta jest ściśle powiązana z wysiłkiem modelowania obliczeniowego w kognitywistyce, gdzie celem modelu obliczeniowego jest odtworzenie ludzkich wyników przy jednoczesnym zapewnieniu mechanizmów wyjaśniających, w jaki sposób ludzie wykonują dane zadanie. W rezultacie sposób, w jaki algorytmy są generowane, wdrażane i testowane, ma wpływ na koncepcje emocji i świadomości, co z kolei będzie wymagało filozoficznego opracowania. Jest oczywiste, że pierwsza postawa jest wystarczająca do osiągnięcia celów badawczych w dziedzinie sztucznej inteligencji (np. zbudowania inteligentnych agentów), jednak druga postawa umożliwi także badaczom sztucznej inteligencji nawiązanie kontaktu z innymi dziedzinami i udostępnienie ich algorytmów i implementacji pod analizę filozoficzną i

psychologiczną. W ten sposób psychologowie mogliby być w stanie opracować nowe projekty eksperymentów, które będą w stanie przetestować przewidywania dokonane na podstawie modeli, a filozofowie mogliby wyostrzyć swoją intuicję dotyczącą tego, do czego te pojęcia mają się odnosić. Historycznie rzecz biorąc, pytania dotyczące tego, czy maszyny mogą odczuwać emocje lub być świadome, pojawiały się w różnych momentach w różnych dziedzinach. W tym miejscu pokrótce dokonamy przeglądu filozoficznych perspektyw dwóch pionierów sztucznej inteligencji i filozofii umysłu – odpowiednio Alana Turinga i Hilary Putnam. Alan Turing w swoim słynnym artykule z 1950 r. „Maszyny obliczeniowe i inteligencja" (Turing 1950) rozważa dziewięć zarzutów wobec swojej „gry w naśladownictwo", która później stała się znana jako „test Turinga".1 Czwarty z nich, „Argument z Świadomość" próbuje odrzucić inteligencję maszyn, wskazując na brak emocji i uczuć w maszynach. W tym miejscu Turing cytuje stwierdzenie profesora Geoffreya Jeffersona. Dopiero gdy maszyna będzie mogła napisać sonet lub skomponować koncert pod wpływem odczuwanych myśli i emocji, a nie przez przypadkowy upadek symboli, moglibyśmy zgodzić się, że maszyna to mózg – to znaczy nie tylko ją napisać, ale wiedzieć, że to napisała . Żaden mechanizm nie mógłby odczuwać (a nie tylko sztucznie sygnalizować, łatwym wynalazkiem) przyjemności ze swoich sukcesów, smutku, gdy jego zastawki się łączą, nie ogrzewać się pochlebstwami, unieszczęśliwiać się swoimi błędami, oczarowywać się seksem, złościć się lub przygnębiać, gdy nie może dostać to, czego chce.

(Cytowanie Turinga 1950)

Turing diagnozuje tę linię argumentacji jako ostatecznie promującą perspektywę solipsystyczną, w której „jedynym sposobem, w jaki można mieć pewność, że maszyna myśli, jest bycie maszyną i odczuwanie, że myśli" (Turing 1950). Wskazuje, że ta sama argumentacja miałaby wówczas zastosowanie również w przypadku ludzi (tj. pewności, że inna osoba ma określone właściwości psychiczne lub jest w określonym stanie psychicznym, można być jedynie wtedy, gdyby była tą inną osobą), problem znany w filozofii jako „problem innych umysłów". Innymi słowy, sprowadza on problem innych umysłów w przypadku maszyn do problemu innych umysłów w przypadku ludzi. Co więcej, zwraca uwagę, że maszyna do pisania sonetów, która udziela przesłuchującemu rozsądnych odpowiedzi na temat własnego sonetu, używając viva voce (a więc prawdopodobnie używając intonacji w sposób po ludzku wiarygodny, łącznie z wyrażaniem emocji), prawdopodobnie nie byłaby postrzegana jako „łatwy wynalazek". Zakłada się, że maszyna potrafiąca porozumiewać się w naturalnym języku mówionym w sposób podobny do ludzkiego, sprawi, że ludzie będą postrzegać ją jako sprawiającą przyjemność, ból itd., w bardzo podobny sposób, w jaki ludzie wnioskują o stanach wewnętrznych innych ludzi na podstawie ich interakcje (np. na podstawie tonu głosu danej osoby). Hilary Putnam szczegółowo ponownie zbadała kwestię, czy maszyny mogą odczuwać uczucia i być świadome. W swoim artykule z 1964 r. „Roboty: maszyny czy sztucznie stworzone życie?" (Putnam 1964), Putnam chce, abyśmy wyobrazili sobie robota Oscara, który jest psychologicznie izomorficzny z człowiekiem – to znaczy ma stany wewnętrzne, które odgrywają tę samą rolę przyczynową, co nasze stany mentalne. Załóżmy, że Oscar ma „wrażenie" czerwieni w tym sensie, wówczas pojawia się pytanie, czy rzeczywiście ma on wrażenie czerwieni, to znaczy, czy Oscar faktycznie coś widzi, czy Oscar czuje, czy Oscar jest przytomny. Podobnie jak Turing, Putnam łączy to pytanie z problemem innych umysłów: „To, czy i w jakich warunkach robot mógłby być świadomy, to kwestia, której nie można omawiać bez od razu nawiązania do tematów omówionych w rozdziałach Umysł–Ciało Problem i problem innych umysłów" (s. 669). Po rozwianiu kilku zarzutów wobec twierdzenia, że Oskar jest świadomy, dochodzi do wniosku, że to pytanie wymaga decyzji, a nie odkrycia. Jeśli mamy podjąć decyzję, wydaje mi się, że lepiej rozszerzyć naszą koncepcję tak, aby roboty były świadome – gdyż „dyskryminacja" na podstawie „miękkości" lub „twardości" części ciała syntetycznego „organizmu" wydaje się głupia jako dyskryminujące traktowanie ludzi ze względu na kolor skóry. Pogląd Turinga i Putnama, że maszyny mogą być w zasadzie świadome, był od tego czasu powtarzany przez różnych

filozofów (np. Lycan 1987). We wszystkich przypadkach zakłada się, że maszyny będą musiały mieć odpowiedni rodzaj wewnętrznej struktury i organizacji poznawczej – odpowiedni typ architektury – aby mogły odczuwać emocje i być świadome (czy wtedy rzeczywiście będą urzeczywistniać emocje i/lub lub być świadomym, będzie zależeć od dodatkowych czynników, jak w przypadku człowieka). Jednak pytanie o właściwy rodzaj architektury, która może realizować emocje i świadomość, jest dokładnie tym, czym próbowały się zająć badania nad sztuczną inteligencją

Emocje w AI

Od początków sztucznej inteligencji badano w różnym stopniu różne formy emocji, pomimo pierwotnego skupienia się sztucznej inteligencji na mechanizmach deliberatywnych, nieemocjonalnych. Jednak ostatnio praca nad emocjami i czynnikami emocjonalnymi stała się znacznie bardziej popularna, między innymi dzięki pracom Aarona Slomana na temat architektur emocjonalnych oraz pracy Rossa Picarda na temat „informatyki afektywnej" , w której podkreślono znaczenie ludzkiego afektu i zbadał, w jaki sposób komputery mogą stać się „świadome afektów" lub emocji. Obecnie jesteśmy świadkami rosnącej liczby społeczności badawczych, które badają aspekty emocji i afektu, od „emocjonalnych" lub „afektywnych" interfejsów użytkownika po „wiarygodne" syntetyczne postacie i realistyczne animowane podmioty posiadające emocje, po emocjonalne lub świadome emocji metody pedagogiczne i agentów instruktażowych, wirtualnych agentów emocjonalnych i robotów . Motywacje poszczególnych kierunków badań i ich szczegółowe cele są oczywiście zupełnie odmienne. Podczas gdy dla niektórych emocje polegają na uczynieniu animowanych postaci bardziej wiarygodnymi (np. poprzez nadanie im emocjonalnego wyrazu twarzy), dla innych rozpoznawanie emocji jest kluczowe, aby system mógł dostosować się do potrzeb użytkownika. Jeszcze inni uważają emocje za integralną część kontroli złożonych czynników i w związku z tym skupiają się na mechanizmach architektonicznych wymaganych w procesach emocjonalnych. Jednak wspólne dla wszystkich tych różnych zachęt do badania emocji jest milczące założenie, że emocje, w takiej czy innej formie, mogą mieć ważne zastosowania w sztucznych czynnikach.

Funkcjonalne role emocji

Jedną z głównych trudności związanych z pojęciami takimi jak emocja (i świadomość) jest to, że nie są one jasno określone, a prawdopodobnie nawet w zasadzie nie można ich jasno określić. Dlatego w psychologii nie ma jasnego pojęcia, czym dokładnie jest emocja (Griffiths 1997), a opisy psychologiczne różnią się znacznie pod względem sposobu indywiduacji emocji (np. na podstawie wyrazu twarzy, wzorców zachowania, obszarów mózgu itp.). Trudności koncepcyjne związane z koncepcjami emocji nie zniechęciły jednak do prób wdrożenia procesów, które przynajmniej przypominają procesy emocjonalne, mimo że badacze sztucznej inteligencji często nie są zgodni co do tego, za co uważają „emocję" i co ich zdaniem oznacza wdrożenie emocje w sztucznych czynnikach . Motywacją wielu badań nad rolą emocji w sztucznych czynnikach była analiza możliwych funkcjonalnych ról emocji w układach naturalnych. Podstawowe założenia są takie, że (1) emocje pełnią role funkcjonalne w architekturach agentów oraz że (2) posiadanie stanów o odpowiednich rolach funkcjonalnych jest wystarczające do posiadania emocji, niezależnie od szczególnej budowy fizycznej agenta. Podczas gdy większość badaczy nauk afektywnych zgodzi się co do (1) (mimo że istnieje również wiele przykładów skutków dysfunkcjonalnego afektu), ich poglądy są odmienne co do (2) – czy wystarczy posiadanie odpowiedniego rodzaju architektury funkcjonalnej do posiadania określonej emocji. Na przykład mogą utrzymywać, że w wielu stanach afektywnych zaangażowane są różne procesy cielesne: Jeśli określone procesy biochemiczne, takie jak wydzielanie określonych hormonów lub zmiany w poszczególnych neuroprzekaźnikach, zostaną uznane za niezbędne lub składające się na afekt, wówczas sztuczne czynniki z definicji nie będą w stanie urzeczywistnić tak rozumianych stanów afektywnych. (Porównaj poglądy niektórych filozofów na temat świadomości lub stanów jakościowych, np. Searle 1992).

Sztuczne czynniki będą jednak w dalszym ciągu zdolne do wywoływania tego samego rodzaju procesów kontrolnych, jakie są realizowane w aktywności neuronowej u zwierząt, ponieważ są one: również z definicji niezależny od budowy fizycznej agenta i może to być wystarczające do celów AI (np. aby sztuczny agent mógł wykonać określone zadanie). Jeżeli natomiast dokładna natura stanów i procesów cielesnych nie odgrywa roli przyczynowej w funkcjonowaniu procesów afektywnych, to do osiągnięcia tych samych efektów można by np. zastosować symulowane układy hormonalne (np. Cañamero 1997).), wówczas sztuczne czynniki będą w stanie urzeczywistnić procesy wpływające, jeśli będą miały odpowiednie warunki wstępne architektoniczne. Niezależnie od tego, jakie stanowisko zajmie się jakościową naturę emocji (tj. w kwestii „jak to jest doświadczyć stanu X"), funkcjonalne aspekty emocji w kontekście systemu kontroli podmiotu można rozpatrywać niezależnie. W szczególności wydaje się, że istnieje dwanaście potencjalnych ról emocji dla sztucznych czynników:

1 Mechanizmy alarmowe – np. szybkie reakcje odruchowe w sytuacjach krytycznych, takie jak procesy lękowe, które przerywają bieżące zachowanie i inicjują reakcję odwrotu, odsuwając agenta od strefy zagrożenia.

2 Wybór działania – np. podjęcie decyzji, co dalej robić na podstawie aktualnego stanu emocjonalnego, np. przejście z eksploracji na poszukiwanie pożywienia w oparciu o potrzeby agenta.

3 Adaptacja – np. krótko- lub długoterminowe zmiany w zachowaniu spowodowane stanami afektywnymi, np. dostosowywanie chodu do nierównego terenu w oparciu o negatywny afekt generowany przez czujniki.

4 Regulacja społeczna – np. wykorzystywanie sygnałów emocjonalnych do osiągnięcia efektów społecznych, np. agresywne okazywanie emocji w celu powstrzymania innego podmiotu od zakłócania czyjejś działalności.

5 Uczenie się – np. wykorzystywanie ocen afektywnych jako szacunków użyteczności w uczeniu się przez wzmacnianie, np. uczenie się przydatności różnych zachowań do osiągania celów w różnych kontekstach.

6 Motywacja – np. przyjmowanie celów jako część mechanizmu emocjonalnego radzenia sobie, na przykład gdy wysoki poziom niepokoju i frustracji prowadzi do przyjęcia celu, jakim jest zwrócenie się o pomoc do ludzkiego przełożonego)

7 Zarządzanie celami – np. tworzenie nowych celów lub zmiana priorytetów istniejących, np. wykorzystanie pozytywnego i negatywnego afektu do modyfikacji szacunków kosztów stosowanych przy kalkulacji oczekiwanej użyteczności celu.

8 Przetwarzanie strategiczne – np. wybór strategii wyszukiwania w oparciu o ogólny stan afektywny, taki jak wykorzystanie pozytywnego i negatywnego wpływu w celu odchylenia algorytmów wyszukiwania od góry do dołu w porównaniu do wyszukiwania od dołu do góry.

9 Kontrola pamięci – np. strategiczne wykorzystanie wpływu afektywnego na dostęp do pamięci i jej odzyskiwanie, a także tempo zanikania elementów pamięci, na przykład wykorzystanie bieżącego stanu afektywnego do uszeregowania elementów pamięci o podobnym afektu jako lepiej dopasowanych.

10 Integracja informacji – np. emocjonalne filtrowanie danych z różnych kanałów informacyjnych lub blokowanie takiej integracji, np. ignorowanie pozytywnie wartościowanych informacji z czujników wizyjnych o wesołej twarzy, gdy informacja akustyczna sugeruje gniewny głos.

11 Koncentracja uwagi – np. wybór danych do przetworzenia w oparciu o ocenę afektywną, taką jak ukierunkowanie poszukiwań wizualnych na korzyść obiektów, których agent bardzo pragnie.

12 Model Ja – np. użycie afektu jako reprezentacji tego, „jak wygląda sytuacja dla podmiotu", na przykład wykorzystanie ogólnej oceny afektywnej różnych elementów systemu kontroli podmiotu jako miary ogólnego nastroju podmiotu i tego, jak on „ czuje".

Chociaż lista ta nie jest wyczerpująca, wskazuje na zróżnicowaną funkcjonalną naturę emocji, od ról architektonicznych po role w regulacji społecznej. Zapewnia także ramy, w których można umiejscowić przeszłe osiągnięcia i przyszłe kierunki badań nad architektonicznymi aspektami afektu

Komunikacyjne a architektoniczne aspekty emocji

Pracę nad emocjami w AI można z grubsza podzielić na dwa wątki (z niewielkim zachodzeniem na siebie): aspekty komunikacyjne i aspekty architektoniczne. Komunikacyjne aspekty emocji dotyczą głównie czwartej roli (regulacja społeczna) i były badane głównie przez społeczności zajmujące się interakcją człowiek-komputer (HCI), a ostatnio także społeczności interakcji człowiek-robot (HRI). Wysiłki skupiają się na rozpoznawaniu emocji, wyrażaniu emocji, a czasami na tym, jak połączyć te dwa elementy, aby poprawić doświadczenia użytkowników z systemem interaktywnym (np. za pośrednictwem interfejsu użytkownika na komputerze lub za pośrednictwem repertuaru sensorycznego i efektorowego robota; Scheutz , Schermerhorn i Kramer 2006). Obie społeczności poczyniły istotne postępy w zrozumieniu rodzajów interakcji emocjonalnych, w jakie angażują się ludzie (np. Brave i Nass 2003) oraz w jaki sposób sprawić, by maszyny je rozpoznawały i sygnalizowały. Drugi główny nurt, architektoniczny aspekt emocji, skupia się na roli i użyteczności emocji w architekturach agentów (takich jak stosowanie ocen emocjonalnych jako szybkich heurystyk w podejmowaniu decyzji), a zatem jest mniej zainteresowany społecznymi komunikacyjnymi aspektami emocji. W tym nurcie próbuje się wykorzystać mechanizmy emocjonalne do poprawy możliwości agenta, a większość prac skupia się na pierwszych pięciu rolach. W szczególności uwagę poświęcono selekcji działań afektywnych lub emocjonalnych, zarówno u agentów symulowanych, jak i agentów robotycznych. Podobnie, sporo pracy poświęcono badaniu użyteczności ewaluacji, które są generowane wewnętrznie i odzwierciedlają pewne aspekty wewnętrznego stanu agenta (a nie zewnętrznych stanów środowiska) dla uczenia się przez wzmacnianie, chociaż większość z tych badań nie nazywa tych ewaluacji „afektywnymi". Jednak zaskakująco niewiele prac skupiło się na badaniu ról od (6) do (12), chociaż istnieją pewne godne uwagi wyjątki . Należy zauważyć, że szczególnie cztery ostatnie role mogą okazać się krytyczne dla systemów refleksyjnych, a co za tym idzie świadomych. Ponieważ, jak zobaczymy , mechanizmy kontroli uwagi, integracji informacji, pamięci roboczej i kontroli dostępu do niej, a także automodelu agenta są uważane za niezbędne składniki rozwoju świadomych maszyn. Istnieje kilka zasadniczych różnic pomiędzy badaniami nad komunikacyjnymi i architektonicznymi aspektami emocji. Co najważniejsze, ten pierwszy nie wymaga tworzenia stanów emocjonalnych w systemie. Na przykład agent nie musi sam być emocjonalny (ani zdolny do odczuwania emocji), aby móc rozpoznać wyraz emocji na ludzkich twarzach. Ci drudzy natomiast muszą twierdzić, że w systemie powstają określone stany emocjonalne. Co więcej, badacze komunikacyjnych aspektów emocji nie potrzebują zadowalającej teorii emocji (tj. teorii stanów emocjonalnych), aby móc stworzyć działające systemy. Możliwość pomiaru zmian w przewodności skóry użytkownika, częstotliwości oddychania itd. i wykorzystanie tych informacji do zmiany poziomu szczegółowości w graficznym interfejsie użytkownika nie oznacza automatycznie, że zmierzono poziom frustracji użytkownika, chociaż w niektórych przypadkach wydaje się to prawdą. W rzeczywistości agent pedagogiczny może poznać ważne fakty na temat swojego użytkownika (np. skuteczność swoich strategii nauczania) w oparciu o takie środki, nie wymagając żadnej reprezentacji procesów emocjonalnych użytkownika ani żadnych samych procesów emocjonalnych. Z drugiej strony

architektury, które twierdzą, że wykorzystują mechanizmy emocjonalne (np. do ustalania priorytetów celów lub odzyskiwania pamięci), będą musiały wykazać, że zaimplementowane mechanizmy rzeczywiście powodują powstanie „stanów emocjonalnych" w jasno określonym sensie. W przeciwnym razie nie ma sensu ani powodu, aby je tak nazywać, chociaż istnieje i zawsze była tendencja w sztucznej inteligencji do przedstawiania uproszczonych programów i robotów AI w taki sposób, jak gdyby uzasadniały one epitety takie jak „emocjonalny", „smutny" „zaskoczony", „przestraszony", „afektywny" i tak dalej, bez żadnej głębokiej teorii uzasadniającej te etykiety. W związku z tym trasa architektoniczna staje przed wyzwaniem dokładnego określenia, co to znaczy „realizować stany emocjonalne" tego rodzaju. Badaczy zajmujących się architekturą emocji w sztucznej inteligencji można dalej podzielić na dwie główne kategorie: tych, którzy próbują modelować jawne, obserwowalne skutki zachowań emocjonalnych (nazywają to modelami okazywania emocji) oraz tych, których celem jest modelowanie wewnętrznych procesów, które powodują na temat zachowań emocjonalnych (nazwij te modele procesów emocji). Większość dotychczasowych prac nad architektonicznymi aspektami emocji w sztucznej inteligencji skupiała się na modelach wyświetlania, których zadaniem jest prawidłowe „mapowanie wejść i wyjść" danego opisu zachowania (np. właściwy rodzaj reakcji emocjonalnej w danym kontekście, np. jako wyraz strachu na twarzy robota, gdy przed nim znajduje się szybko zbliżający się obiekt). W skrajnym przypadku takie mapowanie mogłoby być tak proste, jak to zastosowane w animowanym internetowym agencie zakupowym, który wyświetla zdziwioną twarz, gdy użytkownik próbuje usunąć przedmiot z koszyka. Architektury tego rodzaju można znaleźć w wielu tak zwanych „wiarygodnych agentach", których głównym celem jest nakłonienie ludzkiego obserwatora do myślenia, że agent znajduje się w określonym stanie emocjonalnym. To, czy agent rzeczywiście znajduje się w danym stanie, nie ma znaczenia. W rzeczywistości emocje są tu często przedstawiane jako stany lub wartości „zmiennych emocji" – albo jakościowo, jak sugerują terminy związane z emocjami (np. „szczęśliwy", „boi się" itp.), albo ilościowo, używając wartości liczbowych (np. agent jest „0,4 szczęśliwy", „0,1 przestraszony" itp.). I chociaż niektóre pozwalają agentom przebywać tylko w jednym stanie na raz, inne pozwalają na „mieszanki emocji" (mieszanki jednocześnie obecnych stanów emocjonalnych), w których indywidualne emocje i ich intensywność rozciągają się na wielowymiarową przestrzeń. Należy pamiętać, że cechy te nie powinny oznaczać, że projektowanie architektury było pozbawione motywacji biologicznej. Jest zupełnie odwrotnie: większość (jeśli nie wszystkie) modeli wystaw czerpie inspirację z badań w dziedzinie nauk afektywnych. Jednak ich celem nie jest replikacja żadnych konkretnych danych empirycznych z badań na zwierzętach lub ludziach, ale raczej zbadanie możliwych mechanizmów uzyskania pożądanych, obserwowalnych efektów. Główny problem z modelami okazywania emocji polega na tym, że ostatecznie milczą one na temat roli emocji w architekturach agentów, ponieważ mogą, ale nie muszą, faktycznie wdrażać procesy emocjonalne w celu osiągnięcia pożądanych jawnych zachowań. A nawet jeśli tak się stanie, mogą nam niewiele powiedzieć o roli emocji. Choć bowiem zaimplementowane stany są często oznaczane znanymi terminami, różnią się one znacznie od tych zwykle oznaczanych tymi terminami. Na przykład stan oznaczony jako „niespodzianka" można funkcjonalnie zdefiniować jako wyzwalany przez głośne dźwięki i mający bardzo niewiele wspólnego ze złożonymi procesami leżącymi u podstaw pojęcia „niespodzianki" u ludzi i różnych zwierząt, które obejmują naruszenie przewidywanego wyniku. (W przypadku tak zdefiniowanego stanu bardziej odpowiednią etykietą byłoby „przerażenie"). Natomiast modele procesów mają na celu modelowanie i symulowanie niektórych aspektów procesów emocjonalnych w miarę ich rozwoju. Jak zauważyło wielu psychologów i badaczy sztucznej inteligencji , koncepcje emocji najlepiej scharakteryzować jako oznaczające trwałe procesy kontroli zachowania: akcja i reakcja, dostosowanie i modyfikacja, przewidywanie i kompensacja zachowania w różnych (często społeczne) sytuacje. Często to nie pojedynczy stan wewnętrzny architektury agenta określa, czy agent doświadcza lub okazuje jakieś emocje, ale raczej cała sekwencja takich stanów w połączeniu ze stanami środowiska. Na przykład

„strach” nie odnosi się do charakteru podmiotu w określonym momencie, ale do rozwoju sekwencji zdarzeń, począwszy od postrzegania potencjalnie zagrażających warunków środowiskowych, po reakcję kontroli sprawcy systemu, na reakcję ciała podmiotu, na zmianę percepcji i tak dalej. Modele procesów są zatem znacznie bardziej złożone niż modele poglądowe, ponieważ skupiają się na procesach wewnętrznych (i stanach przetwarzania) związanych z emocjami, zazwyczaj czerpiąc z teorii emocji (psychologicznej, neurologicznej itp.).

Modele procesów emocji

Modele procesów opierają się na różnych składnikach charakterystycznych dla procesów emocjonalnych: komponencie percepcyjnym, który może wywołać proces emocjonalny; składnik trzewny, który wpływa na zmienne homeostatyczne organizmu agenta; komponent poznawczy obejmujący stany przypominające przekonania, a także różne rodzaje deliberacji procesów (np. przekierowanie mechanizmów uwagi, realokacja zasobów przetwarzania, przypomnienie przeszłych epizodów naładowanych emocjonalnie itp.); komponent behawioralny będący reakcją na proces afektu (np. w postaci wyrazu twarzy, gestów lub ruchów ciała itp.); i towarzyszące mu uczucie jakościowe („jak to jest być w stanie S lub go doświadczyć”). Żaden pojedynczy aspekt nie jest konieczny do powstania emocji, ani też żaden pojedynczy aspekt nie jest wystarczający sam w sobie. Jednak większość z nich uważa się za część wielu form ludzkich emocji, które znamy z własnego doświadczenia. Same modele procesów można podzielić na dwie główne klasy w zależności od tego, czy mają na celu wyjaśnienie struktur neurologicznych niskiego poziomu i mechanizmów emocji („modele procesów niskiego poziomu”), czy też mają na celu modelowanie procesów emocjonalnych wyższego poziomu. („modele procesów wysokiego poziomu”). Większość badań nad modelami procesów niskiego poziomu dotyczy warunkowania Pawłowa i jest ukierunkowana na struktury neuronowe i mechanizmy przetwarzania (stąd większość modeli niskiego poziomu to modele „sieci neuronowych”). Modele emocji wyższego poziomu mają na celu uchwycenie większej liczby aspektów poznawczych związanych z procesami afektu i zazwyczaj dotyczą szerszego zakresu afektu (stąd większość modeli wyższego poziomu to modele „symboliczne”). Najszerzej rozwiniętymi ogólnymi modelami niskiego poziomu są modele CogEM Grossberga , które mają na celu pokazanie interakcji pomiędzy emocjonalnymi i nieemocjonalnymi obszarami mózgu (np. ciało migdałowate vs. kora czuciowa lub przedczołowa). Modele CogEM mogą wyjaśniać kilka efektów warunkowania strachem Pawłowa, ale nie zostały bezpośrednio zastosowane do danych empirycznych. Z drugiej strony, specyficzne modele afektu niskiego poziomu mają na celu modelowanie ciała migdałowatego, które pełni kilka funkcji w przetwarzaniu emocji (LeDoux 1996). Na przykład wykazano, że boczne ciało migdałowate bierze udział w warunkowaniu strachem , a wstępny model obliczeniowy uczenia się skojarzeniowego w ciele migdałowatym został opracowany i przetestowany w trzech zadaniach związanych z uczeniem się skojarzeniowym. Co więcej, najnowsze dowody z badań na szczurach sugerują, że ciało migdałowate, w szczególności ciało migdałowate czołowo-skroniowe, integruje informacje sensoryczne i koduje oceny afektywne jako część pamięci strachu . LeDoux i współpracownicy postawili hipotezę o modelu dwuścieżkowym przetwarzania emocjonalnego w ciele migdałowatym, który przetestowali w badaniach słuchowego warunkowania strachowego . Modele te wykorzystano także w badaniach symulowanych zmian chorobowych i z powodzeniem porównano je z danymi z rzeczywistych badań zmian chorobowych na szczurach. Chociaż wszystkie modele niskiego poziomu są modelami sieci neuronowych, modele wyższego poziomu obejmują zarówno podejście koneksjonistyczne, jak i symboliczne. Przykładem podejścia koneksjonistycznego wysokiego poziomu jest model ITERA , który ma na celu badanie, w jaki sposób informacje medialne na temat problemów środowiskowych wpływają na funkcje poznawcze, emocje i zachowanie. Fakty, rodzaje danych wejściowych, emocje i intencje behawioralne są reprezentowane w postaci indywidualnych jednostek neuronowych, które są połączone połączeniami pobudzającymi i hamującymi i konkurują o aktywację.

Większość prób modelowania emocji na wyższych poziomach opiera się jednak na architekturach symbolicznych, na przykład Soar lub ACT . Zwykle skupiają się na modelu OCC (Ortony i in. 1988), który zapewnia „reguły aktualizacji" zmian w stanach emocjonalnych, które można bezpośrednio wdrożyć w systemach opartych na regułach. Obecnie najbardziej zaawansowane implementacje modeli afektu wysokiego poziomu realizowane są w kontekście „wirtualnych ludzi", gdzie można badać użyteczność emocji w sztucznych czynnikach w interakcjach z ludźmi polegających na pełnym zanurzeniu . Jeden konkretny model, model EMA , został również wykorzystany do dalszych teorii psychologicznych, które zakładają, że różne procesy „oceny emocjonalnej i radzenia sobie" są istotnymi częściami ludzkich procesów emocjonalnych. Inne architektury wyższego poziomu próbują wdrożyć różne aspekty psychologicznych teorii emocji; przykłady obejmują model MAMID, którego emocjonalne komponenty „gniew" i „strach" są zgodne z definicją Frijdy , oraz model „niespodzianki" zaproponowany przez Macedo i Cardoso (2001). Istnieje również kilka sugestii koncepcyjnych dotyczących złożonych architektur przypominających człowieka, które wyraźnie uwzględniają ludzkie emocje i poznanie, ale bez zapewnienia konkretnych implementacji proponowanej architektury. Przykłady obejmują model H-CogAff Slomana, maszynę emocji Minsky'ego oraz model trójpoziomowy Normana, Ortony'ego i Revelle'a. Większość modeli emocji została wdrożona i przetestowana w oderwaniu od dowolnego modelu ciała. W związku z tym trudne, jeśli nie niemożliwe, jest zbadanie kluczowych aspektów przetwarzania emocji, które wymagają kontroli ciała i w ten sposób wykraczają poza właściwości funkcjonalne (takie jak skutki warunkowania Pawłowa), które można przetestować w niezależnych modelach (np. poprzez zastosowanie bodźca i pomiar efektu). Podejmowano różne próby włączenia procesów cielesnych do agentów symulowanych i robotycznych. Niektórzy badali wpływ obliczeniowy symulowanych hormonów na kontrolę emocji (Cañamero 1997), podczas gdy inni wdrożyli koneksjonistyczne modele emocji na robotach, w których różne typy emocji są reprezentowane jako jednostki koneksjonistyczne, które rywalizują o aktywację, co z kolei powoduje, że robot wykazuje określone zachowanie . Główna różnica między tymi podejściami a modelami afektu niskiego poziomu i niektórymi modelami wysokiego poziomu opartymi na ocenie polega na tym, że nie próbują one modelować żadnej konkretnej psychologicznej lub neurobiologicznej teorii afektu, np. w celu sprawdzenia lub sfałszowania swoich przewidywań). Zajmują się raczej możliwością zastosowania konkretnego mechanizmu kontrolnego z inżynierskiego punktu widzenia. Główny problem związany z procesowymi modelami afektu wynika bezpośrednio z problemów nękających koncepcje afektu: nie jest jasne, jakiego rodzaju stanu afektywnego jest modelem dany model obliczeniowy. W pewnym sensie modele procesów bez funkcjonalnej charakterystyki realizowanych stanów afektywnych nie są bardziej skuteczne z koncepcyjnego punktu widzenia niż modele poglądowe, które w pierwszej kolejności nie są przeznaczone do implementowania określonych rodzajów stanów afektywnych. Jednakże nawet jeśli z modelu procesu nie można od razu wyciągnąć żadnych koncepcyjnych wniosków, podejście metodologiczne ma ważną zaletę polegającą na próbie wdrożenia hipotetycznych mechanizmów afektu, które przyniosły owoce już w krótkim okresie. Mechanizmy architektoniczne mające na celu umożliwienie powstawania stanów afektywnych można bowiem testować i oceniać jako takie, niezależnie od tego, jakiego rodzaju stany funkcjonalne mogą one wywoływać . Jest to analogiczne do tego, co pragmatycznie wydarzyło się w sztucznej inteligencji w przypadku innych rodzajów architektur, takich jak na przykład architektury wiara-pragnienie-intencja (BDI). Można tutaj postawić tego samego rodzaju pytania pojęciowe dotyczące rzeczywistej natury urzeczywistnionych stanów „przekonania", „pragnienia" i „intencji", podczas gdy mechanizmy architektoniczne rozwiązywania problemów można oceniać niezależnie w różnych dziedzinach pod kątem ich wartości technicznej. Istnieje jednak istotna różnica między podejściami architektonicznymi w dziedzinie rozumowania, rozwiązywania problemów itd. a podejściami architektonicznymi w dziedzinie afektu: to pierwsze często ma dobrze rozwiniętą teorię funkcjonalnego potencjału mechanizmów architektonicznych, podczas gdy podejście architektoniczne ten ostatni nie ma obecnie takiej teorii.

Przeciwnie, badania nad architektonicznymi aspektami afektu są wciąż na etapie przedteoretycznym. Obecny brak dobrze rozwiniętej teorii użyteczności stanów afektywnych w sterowaniu sztucznymi czynnikami nie odbiera jednak faktu, że próby scharakteryzowania i wdrożenia stanów i procesów afektywnych mogą skutkować powstaniem mechanizmów architektonicznych, które mogłyby okazać się przydatne w różnorodne domeny i zastosowania (np. aplikacje, które muszą radzić sobie z poważnymi ograniczeniami zasobów, jak argumentował Scheutz.)

Świadomość maszynowa w AI

W przeciwieństwie do badań nad emocjami, których początki sięgają lat sześćdziesiątych XX wieku, badanie świadomości maszyn w sztucznej inteligencji jest znacznie młodszym przedsięwzięciem, które rozpoczęło się w połowie lat dziewięćdziesiątych i tak naprawdę dopiero zaczyna nabierać rozpędu (chociaż już wcześniej podejmowano pewne próby określenia wymagań dla świadomych maszyn ; patrz np. Angel 1989). Jednym z powodów tego późniejszego rozpoczęcia może być to, że badania nad świadomymi maszynami muszą opierać się na badaniach nad różnymi komponentami funkcjonalnymi niezbędnymi dla świadomości, z których część może być emocjami (sugestia, że emocje i świadomość są nierozerwalnie powiązane). Jednak, co jest nieco zaskakujące, społeczność świadomości maszynowej nie jest podzbiorem społeczności emocji w sztucznej inteligencji ani zbytnio się z nią nie krzyżuje. I chociaż społeczność emocji w sztucznej inteligencji zacieśniła bliskie powiązania z różnymi psychologami i ich teoriami (np. między innymi Andrew Ortony i Craigiem Smithem), społeczność świadomości maszynowej wydaje się być bardziej powiązana z filozofami zainteresowanymi zapewnieniem funkcjonalnego, możliwego do wdrożenia rachunek świadomości. Podobnie jak w przypadku emocji w sztucznej inteligencji, gdzie badacze pracujący nad komunikacyjnym i innymi wymiarami emocji po prostu ignorują pytania o to, czym są emocje i w jaki sposób są realizowane, niektórzy badacze zainteresowani świadomością nie próbują wyjaśniać ludzkiej świadomości. Są raczej zainteresowani „symulowaniem" procesów, które uważają za niezbędne dla świadomości – co (Holland 2003) nazywa „słabą sztuczną świadomością" – lub wykorzystaniem zasad leżących u podstaw ludzkiej świadomości do projektowania lepszych systemów kontroli . Niektórzy jednak interesują się świadomymi maszynami, dlatego podobnie jak badacze procesowych modeli emocji muszą odpowiedzieć sobie na pytanie, co rozumieją przez „świadomość" i, ostatecznie, czego potrzeba, aby to wdrożyć. Jest to oczywiście bardzo trudny problem, zważywszy, że ani filozofowie, ani psychologowie nie są zgodni co do tego, do czego „świadomość" ma się odnosić i co to znaczy być świadomym. (Teorie świadomości rozciągają się od teorii neurologicznych po teorie reprezentacji poznawczych, takie jak różne formy wyższego rzędu , teoria myślenia, która utrzymuje, że myśli i spostrzeżenia stają się świadome ze względu na to, że są celem dalszych myśli lub percepcji.) Podobnie jak w przypadku emocji, badacze sztucznej inteligencji zainteresowani osiągnięciem świadomości w maszynach zaproponowali różne zasady i mechanizmy architektoniczne, które uważają za niezbędne dla świadomych maszyn. Ogólnie rzecz biorąc, propozycje różnią się pod kilkoma względami: (1) stopień, w jakim łączą się z filozofią, psychologią lub neuronauką; (2) zakres, w jakim przedstawiają konkretną architekturę, która może być świadoma, lub szczególne zasady takiej architektury; oraz (3) stopień, w jakim faktycznie zapewniają implementacje swoich architektur lub modeli. Jednakże badacze zgadzają się, że wymagany jest jakiś rodzaj „modelu wewnętrznego", który opiera się na reprezentacjach stanów percepcyjnych agenta i pozwala agentowi symulować lub przewidywać przyszłe zdarzenia i wyniki oraz różne możliwe działania, jakie by to dla niego wyglądało. Badacze nie są jednak zgodni co do dokładnej definicji i rozszerzenia modelu wewnętrznego oraz innych komponentów, z którymi jest on powiązany.

Propozycje architektoniczne

Większość propozycji dotyczących świadomości w sztucznych agentach ma obecnie charakter koncepcyjny i zapewnia zestaw potencjalnie możliwych do wdrożenia zasad (czasami ze wstępnymi implementacjami dla podzbiorów). Na przykład Pentti Haikonen podsumowuje wymagania architektoniczne świadomego systemu w następujący sposób:

(1) Należy opracować odpowiednią metodę reprezentacji informacji. (2) Należy zaprojektować odpowiednie elementy przetwarzania informacji, które pozwolą na manipulację informacją za pomocą wybranej metody reprezentacji. (3) Architektura maszyny, która może pomieścić czujniki, efektory, procesy percepcji, introspekcję i

(3) Należy zaprojektować uziemienie znaczenia, a także przepływ wewnętrznej mowy i wewnętrznych obrazów. (4) Projekt systemu musi uwzględniać także funkcje myślenia i rozumowania, emocje i język.

(Haikonen 2003)

Bardziej formalne podejście przyjmują Aleksander i współpracownicy, którzy wymieniają pięć zasad, określonych jako aksjomaty, które uważa się za wystarczające dla świadomości. Określają pojęcie „świadomy" dla podmiotu i świata w następujący sposób:

Niech A będzie agentem w dostępnym zmysłowo świecie S. Aby A był

świadomy S konieczne jest, aby:

Aksjomat 1 (przedstawienie): A ma stany percepcyjne, które przedstawiają części

z S.

Aksjomat 2 (Wyobraźnia): A ma wewnętrzne stany wyobraźni, które przywołują części S lub wytwarzają wrażenia podobne do S.

Aksjomat 3 (Uwaga): A jest w stanie wybrać, które części S przedstawić lub co sobie wyobrazić.

Aksjomat 4 (Planowanie): A ma środki kontroli nad sekwencjami stanów wyobraźni w celu planowania działań.

Aksjomat 5 (Emocje): A ma dodatkowe stany afektywne, które oceniają zaplanowane działania i determinują późniejsze działanie.

(Aleksander i Dunmall 2003)

Twierdzono, że to połączenie przedstawień zmysłowych, wyobraźni, uwagi i emocji ostatecznie prowadzi do perspektywy pierwszoosobowej („ja" u ludzi). Motywacją aksjomatów nie jest konkretna teoria świadomości, ale duży zbiór indywidualnych odkryć, które wydają się sugerować te zasady jako abstrakcje. Sloman od dłuższego czasu promuje koncepcję „funkcjonalizmu maszyny wirtualnej" jako sposobu wyjaśnienia bogatych procesów wewnętrznych złożonych, przemyślanych i refleksyjnych czynników, które mogą stanowić podstawę introspekcji oraz rozwoju wewnętrznych kategorii i koncepcji, które są niedostępne (nawet poprzez język) na innych agentów, tworząc w ten sposób podstawę perspektywy pierwszoosobowej świadomego agenta (np. Sloman i Chrisley 2003). Jest także kilku innych badaczy, którzy próbują przedstawić funkcjonalne opisy architektoniczne wymagań dotyczących świadomości. Proponowane koncepcje obejmują rozwiązania neuronowe (Shanahan 2005), robotyczne (Kuipers 2005), teorie sterowania (Sanz i in. 2007), oparte na procesach (Manzotti 2003) i inne. Cechą wspólną wszystkich powyższych badaczy jest to, że wdrożyli oni pewne podstawowe modele, które demonstrują części architektury, ale nie kompletny, funkcjonalny, a zatem świadomy system.

Świadomi agenci

Godnym uwagi wyjątkiem wśród badaczy świadomości maszynowej jest praca Franklina i współpracowników (Franklin i in. 1998), którzy próbowali wdrożyć całkowicie świadomego agenta w oparciu o teorię świadomości globalnej przestrzeni roboczej Baarsa. Jest to „teatralny model" świadomości, który wymaga centralnej przestrzeni pracy („sceny"), na której „świadome treści wyłaniają się, gdy jasny reflektor uwagi pada na gracza na scenie pamięci roboczej". Pierwszy funkcjonalny prototyp, „Świadomy" Mattie, był agentem oprogramowania, którego zadaniem było pisanie ogłoszeń o seminariach, komunikowanie się e-mailem z organizatorami seminariów i przypominanie im o spóźnieniach. Drugi prototyp, IDA dla „Inteligentnego agenta dystrybucji", został opracowany dla Marynarki Wojennej Stanów Zjednoczonych w celu ułatwienia procesu przydzielania marynarzy do nowych misji. Obie architektury obejmują mechanizmy „świadomości", obejmujące kontroler reflektora, menedżera transmisji i zbiór kodów uwagi, które rozpoznają nowe lub problematyczne sytuacje, wraz z modułami percepcji, wyboru działań, pamięci skojarzonej, emocji i metapoznania (patrz Franklin 2000). Najnowszym modelem jest kompletna architektura kognitywna o nazwie LIDA (Learning Intelligent Distribution Agent), która dodaje do dotychczasowej architektury różne typy uczenia.

Perspektywy na przyszłość

Badania nad emocjami stały się aktywną interdyscyplinarną dziedziną sztucznej inteligencji, a świadomość maszynowa jest o krok od utworzenia społeczności badawczej zajmującej się projektowaniem świadomych maszyn. Biorąc pod uwagę obecne trajektorie, prawdopodobne jest, że obie społeczności będą się wspólnie rozwijać, zwłaszcza że wspólnota emocji goni za bardziej złożonymi emocjami, takimi jak żal z powodu własnego zachowania lub rozczarowanie postawą innej osoby wobec siebie, które wymagają wielu cech architektonicznych niezbędne dla świadomych maszyn, zgodnie z postulacją społeczności świadomości (reprezentacje własnych percepcji, wewnętrzne skupienie uwagi, wspomnienia przeszłych działań, reprezentacje możliwych przyszłości itp.). Badania w obu obszarach obiecują nie tylko postęp w dziedzinie sztucznej inteligencji, ale także rzucą światło, jeśli nie bezpośrednio na przypadek człowieka, to na przypadek ewentualnych istot emocjonalnych i świadomych, co powinno pomóc nam udoskonalić nasze koncepcje. Co więcej, oba obszary prawdopodobnie przyczynią się do lepszego zrozumienia kompromisów między systemami emocjonalnymi i świadomymi w porównaniu z systemami pozbawionymi jednej lub obu właściwości. Biorąc jednak pod uwagę, że oba przedsięwzięcia są stosunkowo młode, nie powinno być zbyt zaskakujące, że obie dziedziny nie wypracowały zadowalających kryteriów sukcesu ani nie zastanowiły się nad konsekwencjami swojej pracy. „Kryteria sukcesu" mają tu odnosić się do sposobów, które pozwolą nam stwierdzić, czy dana maszyna ma emocje, czy jest świadoma. Prawdopodobnie będzie to obejmować twierdzenia dotyczące architektury funkcjonalnej maszyny i typów obsługiwanych przez nią stanów. Obejmuje to również algorytmy określające, czy dany system faktycznie implementuje architekturę funkcjonalną, ale niestety obecnie brakuje nam również dobrej teorii implementacji (Scheutz 2001a). Idealnie byłoby, gdybyśmy mieli kryteria, które pozwoliłyby ustalić, czy dana maszyna znajduje się w określonym stanie emocjonalnym, czy też jest świadoma. Może to obejmować procedury analogiczne do tych, których używają psychologowie w celu ustalenia, czy dana osoba znajduje się w określonym stanie emocjonalnym lub jest przytomna.

Chociaż szczególna potrzeba stosowania takich kryteriów może nie pojawiać się tak bardzo w przypadku samej sztucznej inteligencji, prawdopodobne jest, że w końcu pojawi się silna presja społeczna, aby rozstrzygnąć te i inne fundamentalne pytania dotyczące natury sztucznych umysłów, zwłaszcza gdy formułowane są twierdzenia dotyczące emocjonalnego i świadome stany maszyn. Tę kwestię dostrzegł Putnam ponad czterdzieści lat temu:

Biorąc pod uwagę stale przyspieszające tempo zmian technologicznych i społecznych, jest całkowicie możliwe, że pewnego dnia roboty będą istnieć i będą twierdziły, że „żyjemy; jesteśmy świadomi!" W takim przypadku to, co dziś jest jedynie uprzedzeniami filozoficznymi o tradycyjnym, antropocentrycznym i mentalistycznym charakterze, najprawdopodobniej przekształciłoby się w konserwatywne postawy polityczne. Ale na szczęście dzisiaj mamy tę przewagę, że możemy omówić ten problem bezinteresownie, co daje nieco większe szanse na znalezienie prawidłowej odpowiedzi.

(Putnam 1964)

O ile Putnam z pewnością miał rację co do konieczności wyjaśnienia pytań dotyczących świadomości maszynowej, o tyle pilność opracowania odpowiedzi na ten problem wyraźnie zmieniła się od chwili, gdy pisał o omawianiu go „bezinteresownie", do chwili obecnej, wraz ze wszystkimi niedawnymi sukcesami w dziedzinie sztucznej inteligencji i systemów autonomicznych. robotyki, a roboty są już rozpowszechniane w społeczeństwie. Dlatego też najwyższy czas, aby badacze i filozofowie sztucznej inteligencji wspólnie zastanowili się nad potencjałem maszyn emocjonalnych i świadomych. Nie chcemy bowiem obudzić się pewnego dnia i odkryć, że to, co traktowaliśmy jako pozbawione emocji, nieświadome artefakty, było w rzeczywistości emocjonalnymi, świadomymi istotami, zniewolonymi i źle traktowanymi przez nas z powodu ignorancji lub uprzedzeń.