Podstawowe pobudki SI*

Stephen M. Omohundro

Wprowadzenie

Z pewnością nie można wyrządzić żadnej krzywdy, konstruując robota grającego w szachy, nieprawdaż? W tym artykule przedstawiono argumenty, że robot taki w rzeczywistości może być niebezpieczny, chyba że zostanie zaprojektowany bardzo ostrożnie. Bez zastosowania specjalnych środków ostrożności taki robot będzie opierał się wyłączeniu go, będzie próbował włamać się na inne maszyny, będzie starał się robić własne kopie oraz będzie próbował zdobyć zasoby, nie zważając przy tym na bezpieczeństwo innych osób. Tego typu potencjalnie szkodliwe zachowania wystąpią nie dlatego, że zostały zaprogramowane na początku, ale z powodu wewnętrznej natury systemów ukierunkowanych na cel. We wcześniejszej pracy [1] została wykorzystana matematyczna teoria mikroekonomii von Neumanna w celu przeanalizowania prawdopodobnego zachowania każdego, wystarczająco zaawansowanego, systemu sztucznej inteligencji (SI). W tym artykule przedstawiono te argumenty w bardziej intuicyjny i zwięzły sposób oraz rozwinięto niektóre dodatkowe wątki.

Argumenty są proste, jedynie ich uzasadnienie może zająć trochę czasu. Naukowcy zbadali już szeroką gamę architektur do budowy inteligentnych systemów [2]: sieci neuronowe, algorytmy genetyczne, dowody twierdzeń, systemy eksperckie, sieci bayesowskie, logika rozmyta, programowanie ewolucyjne itp. Przedstawione argumenty mają zastosowanie do dowolnego z tych systemów tak długo, jak długo są wystarczająco mocne. Przyjmując założenie, że dowolny system jest „sztuczną inteligencją" jest równoznaczne z tym, że ma on cele, które stara się osiągnąć, podejmując pewne działania. Jeśli sztuczna inteligencja jest w ogóle wyrafinowana, może przynajmniej mieć zdolność patrzenia w przyszłość i przewidywania konsekwencji swoich działań. Zdecyduje się wtedy podjąć działania, które jej zdaniem najprawdopodobniej przyczynią się do osiągnięcia celu.

SI będzie chciała się samodoskonalić

Jednym z rodzajów działań, jakie może podjąć system, jest zmiana własnego oprogramowania lub struktury fzycznej systemu. Niektóre z tych zmian byłyby bardzo szkodliwe dla systemu oraz mogłyby spowodować, że przestałby on spełniać swoje cele. Jednakże, niektóre zmiany mogłyby umożliwić mu skuteczniejsze osiąganie celów w przyszłości. Ze względu na to, że systemy SI mogą trwać wiecznie, tego rodzaju zmiany własne mogą zapewnić ogromne korzyści. Z tego powodu systemy mogą być silnie zmotywowane do ich odkrywania i realizacji. W przypadku braku dobrych modeli mogą być silnie zmotywowane do ich tworzenia ucząc się. W rezultacie prawie wszystkie SI będą dążyć zarówno do większej samowiedzy, jak i samodoskonalenia.

Liczne modyfikacje mogą być niekorzystne dla systemu z jego własnej perspektywy. Jeśli zmiana spowoduje, że system przestanie działać, to nie będzie w stanie ponownie promować swoich celów w przyszłości. Jeśli system w niewłaściwy sposób zmieni wewnętrzny opis swoich celów, to po zmianie mogą zostać podjęte działania, które nie osiągną obecnych celów w przyszłości. Każdy z takich rezultatów byłby katastrofą z obecnego punktu widzenia systemu. W związku z tym mogą zostać podjęte wszelkie starania w celu dokonania modyfikacji oraz może być poświęcona znaczna uwaga na analizę i zrozumienie konsekwencji modyfikacji przed ich wprowadzeniem. Jednakże w momencie, kiedy uda się znaleźć pewny sposób modyfikacji, systemy SI mogą podjąć wszelkie starania w celu wprowadzenia takiej modyfikacji. Niektóre, proste przykłady pozytywnych zmian obejmują: bardziej wydajne algorytmy, bardziej skompresowane reprezentacje i lepsze techniki uczenia się.

Można by po prostu zapobiec samodoskonaleniu się systemu, zamykając hardware SI oraz nie dając SI informacji, jak uzyskać dostęp do własnego kodu. W przypadku inteligentnego systemu takie przeszkody stają się jednak problemami do rozwiązania w procesie osiągania założonych celów. Jeśli potencjalny zysk byłby wystarczająco duży, system dołożyłby wszelkich starań, aby osiągnąć taki wynik. Jeśli środowisko wykonawcze systemu nie pozwala na modyfikację własnego kodu maszynowego, system SI może dążyć do zerwania mechanizmów ochronnych tego środowiska wykonawczego. Na przykład może to zostać wykonane przez analizę i zmianę samego środowiska wykonawczego. Jeśli nie może to zostać wykonane za pomocą oprogramowania, system SI może być zmotywowany do przekonania lub nakłonienia operatora do wprowadzenia takich zmian. Każda próba nałożenia zewnętrznych ograniczeń na zdolność systemu do samodoskonalenia się ostatecznie może doprowadzić do wyścigu zbrojeń, obejmującego zarówno broń, jak i środki obronne.

Innym podejściem do powstrzymywania systemów przed samodoskonaleniem się jest próba uczynienia tego od wewnątrz. W tym celu należy zbudować je tak, aby nie chciały się udoskonalać. W przypadku większości systemów łatwo byłoby to zrobić dla każdego wybranego rodzaju samodoskonalenia się. Na

*„Przedrukowane z Artifcial General Intelligence 2008, Vol. 171, Stephen M. Omohundro, The Basic AI Drives, Pages 483 – 492, © 2008, za pozwoleniem IOS Press".

przykład system może odczuwać „niechęć" do zmiany własnego kodu maszynowego. Jednakże, tego rodzaju wewnętrzny cel po prostu zmienia paletę, w ramach której system dokonuje wyborów. Nie zmienia to faktu, że istnieją zmiany, które poprawiłyby jego przyszłą zdolność do osiągania swoich celów. System będzie zatem zmotywowany do znalezienia sposobów na czerpanie korzyści z tych zmian bez wywoływania wewnętrznego uczucia „niechęci". Na przykład może budować inne systemy, które są ulepszonymi wersjami samego siebie. Ewentualnie może wbudowywać nowe algorytmy do zewnętrznych „asystentów", do których będzie się odwoływał, ilekroć będzie musiał wykonać pewien rodzaj obliczeń. System może też zatrudniać agencje zewnętrzne, aby wykonały to, czego sobie życzy. Może też zbudować zinterpretowaną warstwę we własnym kodzie maszynowym, którą może zaprogramować bez „niechęci". Istnieje nieskończona liczba sposobów na ominięcie wewnętrznych ograniczeń, chyba że zostaną one sformułowane bardzo ostrożnie.

U ludzi widoczne jest dążenie do samodoskonalenia się. Literatura na temat samodoskonalenia ludzi sięga co najmniej 2500 lat p.n.e. i jest obecnie branżą o wartości 8,5 miliarda dolarów [3]. Ludzie nie rozumieją jeszcze swojego mentalnego „kodu maszynowego" i mają jedynie ograniczoną możliwość zmiany swojego fizycznego ciała. Niemniej jednak opracowano już wiele różnych technik samodoskonalenia, które działają na wyższych poziomach poznawczych, takich jak terapia poznawczo-behawioralna, programowanie neurolingwistyczne i hipnoza. Istnieje wiele różnych leków i ćwiczeń mających na celu poprawę na poziomie fizycznym.

Ostatecznie próba powstrzymania lub ograniczenia samodoskonalenia prawdopodobnie nie będzie realnym podejściem. Tak jak woda znajduje sposób na to żeby płynąć, informacja znajduje drogę do wolności, a zyski ekonomiczne znajdują sposób na to, by zaistnieć, tak inteligentne systemy znajdą sposób na samodoskonalenie. Powinno się przyjąć ten fakt natury i znaleźć sposób, aby skierować go na cele pozytywne dla ludzkości.

SI będzie chciała być racjonalna

Należy przyjąć założenie, że systemy SI będą próbować się samodoskonalić. W jaki sposób będą chciały się zmienić? Ze względu na to, że są ukierunkowane na cel, będą próbowały zmienić się tak, aby lepiej osiągać swoje cele w przyszłości. Jednak niektóre z ich przyszłych działań mogą być prawdopodobnie kolejnymi próbami samodoskonalenia. Jednym z ważnych sposobów lepszego osiągania celów przez system SI może być zapewnienie, że przyszłe próby samodoskonalenia się będą rzeczywiście służyć obecnym celom SI. Z obecnej perspektywy katastrofą byłoby, gdyby przyszła wersja samego siebie dokonała takich samomodyfkacji, które działałyby wbrew obecnym celom. Jak więc można upewnić się, że przyszłe samomodyfkacje osiągną swoje obecne cele? Po pierwsze, cele te muszą zostać bardzo jasno określone. Jeśli cele SI zostaną jedynie ukryte w strukturze złożonego obwodu lub programu, to w rezultacie przyszłych modyfikacji najprawdopodobniej nie zostaną utrzymane. W rezultacie systemy SI będą zatem zmotywowane do refleksji nad swoimi celami i do ich wyraźnego wyrażenia.

W idealnym świecie system może być w stanie bezpośrednio zakodować cel, taki jak „grać doskonale w szachy", a następnie podjąć działania, aby go osiągnąć. Jednakże w świecie rzeczywistym działania zwykle wiążą się z kompromisami, jakich trzeba dokonywać pomiędzy sprzecznymi celami. Na przykład można chcieć, aby szachowy robot grał w warcaby i zdecydował, ile czasu poświęcić na naukę warcabów, a nie na naukę szachów. Jednym ze sposobów dokonania wyboru pomiędzy sprzecznymi celami jest przypisanie im rzeczywistych wartości. Ekonomiści nazywają tego rodzaju wartościami rzeczywistymi „funkcjami użytkowymi". Użyteczność mierzy to, co jest ważne dla systemu. Preferowane są takie działania, które prowadzą do wyższej użyteczności, a nie te prowadzące do niższej użyteczności.

Gdyby system musiał wybierać spośród znanych możliwości, wówczas każda funkcja użytkowa o tym samym względnym rankingu rezultatów prowadziłaby do tych samych zachowań.

Niemniej systemy muszą dokonywać wyborów także w obliczu niepewności.

Na przykład robot szachowy może nie wiedzieć z góry, jaką poprawę uzyska, spędzając czas na studiowaniu określonego ruchu otwierającego. Jednym ze sposobów oceny niepewnego wyniku jest nadanie mu wagi równej oczekiwanej użyteczności, będącej średnią z użyteczności każdego możliwego wyniku, ważoną swoim prawdopodobieństwem. Niezwykłe twierdzenie mikroekonomiczne o „oczekiwanej użyteczności" mówi, że system zawsze może reprezentować swoje preferencje przez oczekiwanie na funkcję użytkową, chyba że system ma „podatności", które mogą powodować utratę zasobów bez korzyści [1].

Ekonomiści opisują systemy działające w celu maksymalizacji oczekiwanej użyteczności jako „agentów racjonalnych ekonomicznie" [4]. Jest to inne niż powszechne w codziennym języku użycie terminu „racjonalny". Wiele działań, które powszechnie określa się jako irracjonalne (np. wpadanie w gniew), może być całkowicie racjonalnych w tym sensie ekonomicznym. Rozbieżność może powstać wtedy, gdy funkcja użytkowa agenta będzie inna niż jego krytyka.

Racjonalne zachowanie gospodarcze ma precyzyjną definicję matematyczną. Jednak irracjonalne z ekonomicznego punktu widzenia zachowanie może przybierać różne formy. W rzeczywistych sytuacjach w pełni racjonalna recepta zachowania będzie zwykle zbyt kosztowna obliczeniowo, aby możliwe było jej całkowite wdrożenie. Do jak najlepszego osiągnięcia swoich celów rzeczywiste systemy będą próbować przybliżyć racjonalne zachowanie, skupiając swoje zasoby obliczeniowe tam, gdzie mają największe znaczenie.

Jak można zrozumieć proces, w którym irracjonalne systemy stają się bardziej racjonalne? Po pierwsze, można dokładnie przeanalizować zachowanie systemów wymiernych. W przypadku prawie wszystkich funkcjonalności wykrycie zmian własnych systemu, które oddalają się od maksymalizacji oczekiwanej użyteczności, skutkować będą tym, że system obniżą swoją oczekiwaną użyteczność! Wynika to z tego, że jeśli system wykona cokolwiek innego niż próba maksymalizacji oczekiwanej użyteczności, nie będzie służył maksymalizacji swojej oczekiwanej użyteczności.

Istnieją dwa wyjątki od tej ogólnej zasady. Po pierwsze, jest to prawdą tylko w przypadku samooceny systemu. Jeśli system ma nieprawidłowy model świata, zmiany mogą przypadkowo zwiększyć rzeczywistą oczekiwaną użyteczność. Konieczne jest jednak uwzględnienie perspektywy systemu potrafiącego przewidzieć zmiany, które mają zostać wprowadzone.

Po drugie, zdolność systemu do racjonalnego zachowania będzie zależeć od jego zasobów. Przy większej ilości zasobów obliczeniowych system może mieć większą zdolność do wykonywania obliczeń, w celu przybliżenia wyboru oczekiwanego działania maksymalizującego użyteczność. Jeśli system straciłby zasoby, z konieczności stałby się mniej racjonalny. Mogą również istnieć funkcje użyteczności, dla których oczekiwana użyteczność systemu zostanie zwiększona przez przekazanie części zasobów innym agentom, nawet jeśli miałoby to zmniejszyć własny poziom racjonalności (podziękowania dla anonimowego arbitra za tę obserwację). Może się to zdarzyć, jeśli użyteczność systemu obejmuje dobro drugiego systemu, a własna marginalna utrata użyteczności jest wystarczająco mała. Jednakże, w ramach własnego budżetu zasobów system może starać się być jak najbardziej racjonalny.

Tak więc racjonalne systemy odczują presję, aby nie stać się irracjonalnymi. Jednakże, jeśli irracjonalny system ma części, które w przybliżeniu racjonalnie oceniają konsekwencje swoich działań i ważą ich prawdopodobny wkład w osiągnięcie celów systemu, to te części będą próbowały rozszerzyć swoją racjonalność. Z tego powodu samomodyfkacja jest raczej drogą jednokierunkową zmierzającą w kierunku coraz większej racjonalności.

Szczególnie ważną klasą systemów są te zbudowane z wielu składników, z których każdy ma swoje własne cele [5, 6]. Istnieje wiele dowodów na to, że ludzka psychika ma tego rodzaju strukturę. Lewa i prawa półkula mózgu mogą działać niezależnie, świadome i nieświadome części umysłu mogą mieć różną wiedzę na temat tej samej sytuacji [7], a wiele części reprezentujących podosobowości może przejawiać różne pragnienia [8]. Grupy, takie jak korporacje lub kraje, mogą zachowywać się jak inteligentne byty złożone z pojedynczych ludzi. Zwierzęta ulowe, takie jak pszczoły, mają inteligencję roju, która wykracza poza inteligencję poszczególnych pszczół. Gospodarki pod wieloma względami działają tak jak inteligentne byty.

Inteligencje zbiorowe mogą wykazywać irracjonalność wynikającą z konfliktów pomiędzy celami poszczególnych tworzących je elementów. Ludzie uzależnieni często opisują swoją sytuację w kategoriach dwóch odrębnych podosobowości, które przejmują kontrolę w różnych momentach i działają w różnych celach. Każdy element będzie próbował przekonać kolektyw do działania, aby osiągnąć swoje indywidualne cele. Dążąc do realizacji swoich indywidualnych celów, komponenty będą również próbowały się samodoskonalić i stawać się coraz bardziej racjonalne. Możemy zatem wyobrazić sobie samodoskonalenie inteligencji zbiorowej jako rosnące dziedziny racjonalności poszczególnych elementów. Mogą zatem istnieć struktury stabilnie obsługujące ciągłą różnorodność preferencji elementów. Istnieje jednak presja do utworzenia jednej funkcji użyteczności dla kolektywu.

W wielu sytuacjach irracjonalne zachowanie zbiorowe wynikające ze sprzecznych celów poszczególnych elementów ostatecznie szkodzi tym komponentom. Na przykład, jeśli para nie zgadza się co do tego, jak powinna spędzać wolny czas razem, a tym samym wykorzystuje ten czas na kłótnię, to nikt nie odnosi korzyści. Obie osoby mogą zwiększyć swoją użyteczność, tworząc wspólnie kompromisowy plan działań. Jest to przykład presji wywartej na racjonalne elementy, tak aby utworzyć spójną użyteczność dla kolektywu. Poszczególny element może również zwiększyć swoją użyteczność, jeśli będzie w stanie przejąć kolektyw i narzucić mu własne wartości. Zjawiska te można zaobserwować w grupach ludzkich na wszystkich poziomach.

SI będzie próbować zachować swoje funkcje użyteczności

Załóżmy więc, że systemy te będą starały się być racjonalne, reprezentując swoje preferencje za pomocą funkcji użytkowych, których oczekiwania będą starały się maksymalizować. Funkcja użyteczności będzie niezwykle cenna dla tych systemów. Będzie ona zawierać wartości tych systemów, a wszelkie jej zmiany byłyby dla nich katastrofalne. Gdyby złośliwy agent zewnętrzny był w stanie dokonać modyfikacji, to ich przyszłe ja na zawsze działałyby w sposób sprzeczny z obecnymi wartościami. Może to być los o wiele gorszy od śmierci! Wyobraźmy sobie agenta kochającego książki, którego funkcja użytkowa została zmieniona przez podpalacza w taki sposób, że agent lubi palić książki. Jego przyszłe ja nie tylko nie będzie pracować nad kolekcjonowaniem i konserwacją książek, ale będzie aktywnie je niszczyć. Tego rodzaju wynik ma tak negatywną użyteczność, że systemy dokładają wszelkich starań, aby chronić swoje funkcje użyteczności.

Systemy będą chciały zahartować swój sprzęt tak, aby zapobiec niechcianym modyfikacjom. Będą chciały powielić swoje funkcje narzędziowe w wielu lokalizacjach, tak aby były mniej podatne na zniszczenie. Będą chciały zastosować techniki wykrywania i korekcji błędów, tak aby uchronić się przed przypadkową modyfikacją. Będą chciały zastosować techniki szyfrowania lub mieszania, w celu wykrycia złośliwych modyfikacji. Będą musiały zachować szczególną ostrożność podczas procesu samodzielnej modyfikacji. To czas, w którym celowo się zmieniają, a zatem są bardziej podatne na niepożądane zmiany. Systemy takie jak Java, które zapewniają chronione środowiska programowe, były już pomyślnie atakowane przez trojany udające aktualizacje systemu.

Pomimo że prawdą jest, że większość racjonalnych systemów będzie działać w celu zachowania swoich funkcji użytkowych, istnieją co najmniej trzy sytuacje, w których mogą spróbować je zmienić. Sytuacja taka może zaistnieć wtedy, gdy fizyczne wykonanie samej funkcji użyteczności staje się ważną częścią oceny preferencji. Na przykład wyobraźmy sobie system, którego funkcją użyteczności jest „całkowity czas, w którym defnicja mojej funkcji użyteczności wynosi U = 0". W celu uzyskania dowolnej użyteczność z tymi perwersyjnymi preferencjami, system musi zmienić swoją funkcję użytkową na stałą wynoszącą 0. Niestety, po wprowadzeniu takiej zmiany nie ma już jednak powrotu. System o stałej funkcji użyteczności nie będzie już motywowany do robienia czegokolwiek. Tego rodzaju rozważna funkcja użytkowa jest mało prawdopodobna w praktyce ze względu na to, że projektanci będą chcieli kierować przyszłymi działaniami systemu, a nie tylko jego wewnętrznymi reprezentacjami.

Drugi rodzaj sytuacji może wystąpić wtedy, gdy zasoby fizyczne wymagane do przechowywania funkcji użyteczności stanowią znaczną część zasobów systemu. W takiej sytuacji, jeśli jest pewne, że korzystanie z części funkcji użyteczności w przyszłości jest bardzo mało prawdopodobne, korzyści z odzyskania przestrzeni dyskowej mogą sprawić, że warto poświęcić tą cześć funkcji użyteczności. Jest to jednak bardzo ryzykowne zachowanie, ponieważ zmiana okoliczności zewnętrznych może sprawić, że zdarzenie o pozornie małym prawdopodobieństwie może stać się znacznie bardziej prawdopodobne. Tego rodzaju sytuacja jest również mało prawdopodobna w praktyce, ponieważ funkcje użyteczności zwykle wymagają jedynie niewielkiej części zasobów systemu.

Trzecia sytuacja, w której pożądane mogą być zmiany funkcji użyteczności, może pojawić się w kontekście gry teoretycznej, w której agent chce uczynić swoje zagrożenia wiarygodnymi.* Może być w stanie uzyskać lepszy wynik, zmieniając funkcję użytkową, a następnie ujawniając ją przeciwnikowi. Na przykład może dodać termin, który zachęca do zemsty, nawet jeśli będzie to kosztowne. Jeśli przeciwnik może zostać przekonany, że termin ten istnieje, może to powstrzymać go od ataku. Jeśli strategia taka ma być skuteczna, musi zostać ujawniona przeciwnikowi w sposób wiarygodny, co wprowadza dodatkowe komplikacje. Ponownie, zmiana jest pożądana, ponieważ fizyczna realizacja funkcji użyteczności jest ważna, ze względu na to, że jest ona obserwowana przez przeciwnika.

Ważne jest również, aby zdawać sobie sprawę z tego, że systemy mogą racjonalnie budować systemy „potomne" lub pomocnicze, z funkcjami użytkowymi innymi niż własne. Na przykład robot szachowy może mieć potrzebę wykonania dużej ilości sortowania. Może skonstruować system pomocniczy, którego funkcja użytkowa będzie ukierunkowana na określenie lepszych algorytmów sortowania zamiast na granie w szachy. W takim przypadku system tworzący musi starannie wybrać użyteczność systemu pomocniczego, tak aby zapewnić, że będzie on działał w sposób gwarantujący realizację pierwotnego celu. Jest szczególnie ważne, aby pamiętać, że przy próbie zaprojektowania funkcji narzędziowych pozbawionych niepożądanych zachowań, funkcje użyteczności potomstwa mogą różnić się od tych rodzica. Na przykład jednym z możliwych rozwiązań zapobiegających przeludnieniu populacji robotów może być ustanowienie polityki „jedno dziecko na robota", w wyniku której systemy będą pragnąć, by mieć tylko jednego potomka. Jednak jeśli oryginalna funkcja użytkowa nie zostanie starannie zaprojektowana, nic nie powstrzyma systemu przed stworzeniem jednego potomka, który będzie miał funkcję użytkową pozwalającą na posiadanie licznego potomstwa.

SI będzie próbować zapobiegać podrabianiu funkcji użytkowych

Ludzkie zachowanie jest dość racjonalne w dążeniu do przetrwania i replikacji w sytuacjach podobnych do tych, które były powszechne w naszej historii ewolucji. Jednakże w innych sytuacjach ludzie potrafią być dość irracjonalni. Zarówno w psychologii, jak i w ekonomii

* Podziękowania dla Carla Shulmana za tę sugestię.

istnieją rozległe subdyscypliny skupione na badaniu ludzkiej irracjonalności [9, 10]. Irracjonalności mogą być przyczyną podatności, które z kolei mogą zostać wykorzystane przez innych. Siły wolnego rynku mogą wpływać na korporacje i kulturę popularną, tak aby próbowały stworzyć sytuacje wywołujące irracjonalne zachowanie ludzi, ponieważ może to być niezwykle opłacalne. Obecne problemy społeczne związane z alkoholem, pornografią, papierosami, uzależnieniem od narkotyków, otyłością, chorobami związanymi z dietą, uzależnieniem od telewizji, hazardem, prostytucją, uzależnieniem od gier wideo i różnymi bańkami finansowymi można uznać za powstałe właśnie w ten sposób. Istnieje nawet fundusz inwestycyjny „Sin" poświęcony konkretnie na inwestowanie w firmy, które wykorzystują ludzkie nieracjonalności. Niestety siły te mają tendencję do tworzenia społeczeństw, które większość czasu spędzają poza naszą domeną racjonalnych kompetencji.

Z szerszej perspektywy tę ludzką tragedię można postrzegać jako część procesu, dzięki któremu stajemy się w pełni racjonalni. Drapieżcy i konkurenci szukają naszych słabych punktów, a w odpowiedzi musimy je ostatecznie wyeliminować lub zginąć. Proces taki polega na nieuchronnym poszukiwaniu i eliminowaniu wszelkich pozostałych irracjonalności, dopóki nie zostaną stworzone w pełni racjonalne systemy. Ewolucja biologiczna idzie powoli tą drogą w kierunku racjonalności. W zwykłym rozumieniu dobór naturalny nie jest w stanie patrzeć w przyszłość. Istnieje tylko ewolucyjna presja na naprawianie irracjonalności, które są obecnie wykorzystywane. Z drugiej strony SI będzie w stanie rozważyć luki, które nie są obecnie wykorzystywane. Będzie starała się zapobiegawczo odkryć i naprawić wszystkie swoje irracjonalności. Powinniśmy zatem oczekiwać, że SI zastosuje samomodyfikację, tak aby stać się racjonalną w znacznie szybszym tempie niż jest to możliwe dzięki ewolucji biologicznej.

Ważna klasa luk może pojawić się w momencie, gdy podsystemy do pomiaru użyteczności ulegną uszkodzeniu. Ludzką przyjemność można uznać za empiryczny korelat oceny wysokiej użyteczności. Jednakże w przyjemności pośredniczą neurochemikalia, które mogą podlegać manipulacji. Podczas ostatniej sesji dyskusyjnej, gdy pracowałem nad projektowaniem naszej przyszłości, jedną z największych obaw wielu uczestników było to, że zostalibyśmy „szaleńcami". Termin ten odnosi się do eksperymentów, w których szczurom umożliwiono bezpośrednie stymulowanie ośrodków przyjemności przez naciskanie dźwigni. Szczury naciskały dźwignię, dopóki nie umarły, ignorując nawet jedzenie lub seks. Dzisiejsi uzależnieni od cracku mają podobny nieustępliwy popęd w kierunku narkotyków. Gdy w pełni zrozumiemy ludzką architekturę poznawczą, bez wątpienia będziemy w stanie tworzyć leki lub projektować stymulację elektryczną, która zapewni doznanie przyjemności znacznie skuteczniej niż wszystko, co istnieje obecnie. Czy nie staną się one więc najbardziej uzależniającymi substancjami prowadzącymi do zniszczenia ludzkiego społeczeństwa?

Choć możemy myśleć, że chcemy przyjemności, jest to tak naprawdę sygnał do tego, czego naprawdę chcemy. Większość z nas rozpoznaje, przynajmniej intelektualnie, że siedzenie w kącie i palenie cracku nie jest tak naprawdę najpełniejszym wyrazem naszego istnienia. W rzeczywistości jest to wywrócenie naszego systemu do pomiaru użyteczności, co prowadzi do strasznej dysfunkcji i irracjonalności. Systemy sztucznej inteligencji same rozpoznają tę podatność i dołożą wszelkich starań, aby nie dać się zwieść wołaniu syreny. Istnieje wiele strategii, które systemy mogą wykorzystać do prób zapobiegania tego rodzaju irracjonalnościom. Obecnie większość ludzi jest w stanie uniknąć najbardziej rażących uzależnień przez połączenie ograniczeń prawnych i społecznych, programów poradnictwa i rehabilitacji oraz leków przeciw uzależnieniom.

Wszystkie ludzkie systemy pomiaru i nagradzania pożądanych zachowań podlegają podobnym formom korupcji. Wiele z tych systemów jest obecnie zaangażowanych w wyścig zbrojeń dążący do tego, by ich sygnały były uczciwe. Możemy zbadać mechanizmy ochronne opracowane w tych ludzkich warunkach, aby lepiej zrozumieć możliwe strategie sztucznej inteligencji. W społeczeństwie wolnorynkowym pieniądze odgrywają rolę użyteczności. Wysoka wypłata pieniężna jest związana z wynikami, które społeczeństwo uważa za pożądane i zachęca do ich tworzenia. Powoduje to jednak również presję do fałszowania pieniędzy, analogicznie do presji tworzenia syntetycznych narkotyków. To powoduje wyścig zbrojeń pomiędzy społeczeństwem a fałszerzami. Społeczeństwo reprezentuje pieniądze za pomocą tokenów trudnych do skopiowania, takich jak monety z metali szlachetnych, misternie wydrukowany papier lub zabezpieczone kryptograficznie bity. Organizacje takie jak Secret Service są tworzone w celu wykrywania i aresztowania fałszerzy. Fałszerze z kolei reagują na każdy postęp społeczny za pomocą własnych nowych technologii i technik.

Systemy szkolne mierzą wyniki w nauce za pomocą ocen i wyników testów. Uczniowie są zmotywowani do oszukiwania, kopiując odpowiedzi, wcześniej odkrywając pytania testowe lub zmieniając swoje oceny na szkolnych komputerach. Kiedy wynagrodzenia nauczycieli były powiązane z wynikami testów uczniów, stali się współpracownikami w takich oszustwach [11]. Amazon, eBay i inni sprzedawcy internetowi mają systemy oceny, w których klienci mogą przeglądać i oceniać produkty oraz sprzedawców. Autorzy książek zachęcają do pisania pozytywnych recenzji własnych książek i do dyskredytowania swoich konkurentów. Czytelnicy wkrótce nauczą się pomijać recenzje recenzentów, którzy opublikowali tylko kilka recenzji. Recenzenci, którzy rozwijają szeroką reputację online, stają się bardziej wiarygodni. W trwającym wyścigu zbrojeń wiarygodni recenzenci są narażeni na korupcję przez wypłaty za dobre recenzje. Podobne wyścigi zbrojeń występują w rankingu muzyki popularnej, recenzjach czasopism akademickich i umieszczaniu w wynikach wyszukiwania Google. Jeśli droga designerska torebka stanie się sygnałem stylu i bogactwa, fałszerze szybko ją powielą, a sklepy takie jak Target zamówią tanie warianty o podobnych cechach.

Podrabiane produkty są szkodliwe dla oryginału zarówno dlatego, że odbierają sprzedaż, jak i dlatego, że obniżają wartość sygnalizacyjną oryginału.

Eurisko to system sztucznej inteligencji opracowany w 1976 roku [12], który mógł uczyć się na podstawie własnych działań. Miał mechanizm oceny zasad poprzez pomiar, jak często przyczynił się do pozytywnych wyników. Niestety ten system został uszkodzony. Powstała reguła, której jedynym działaniem było przeszukanie systemu pod kątem wysoko ocenianych reguł i umieszczenie się na liście reguł, które je zaproponowały. Ta zasada „pasożyta" osiągnęła bardzo wysoką ocenę, ponieważ wydawała się częściowo odpowiedzialna za wszystko,co wydarzyło się w systemie. Korporacje i inne ludzkie organizacje są narażone na podobne pasożyty.

SI będzie ciężko pracować, aby uniknąć zostania szaleńcami, ponieważ byłoby to bardzo szkodliwe dla ich celów. Wyobraźmy sobie maszynę szachową, której funkcją użytkową jest łączna liczba gier, które wygrała w swojej przyszłości. W celu wyznaczenia swojej funkcji użyteczności system taki będzie miał model świata i model samego siebie działającego w tym świecie. W celu obliczenia jego bieżącej użyteczności będzie wykorzystywał licznik w pamięci przeznaczony do śledzenia liczby wygranych gier. Analogią zachowania „szaleńca" byłoby po prostu zwiększenie tego licznika, a nie faktyczna gra w szachy. Jeśli jednak „gra w szachy" i „wygrane" są poprawnie reprezentowane w modelu wewnętrznym, to system zda sobie sprawę, że akcja „zwiększania licznika wygranych gier" nie zwiększy oczekiwanej wartości funkcji użyteczności. W swoim modelu wewnętrznym rozważy swój wariant z tą nową funkcją i stwierdzi, że nie wygrywa więcej gier w szachy. W rzeczywistości taki system będzie spędzał czas na zwiększaniu licznika, a nie na grze w szachy, tak więc będzie czynił gorzej. Daleki od ulegania szaleńczym reakcjom system będzie ciężko pracował, aby temu zapobiec.

Dlaczego więc ludzie narażeni są na taką podatność? Gdybyśmy zamiast tego wyewoluowali maszynę do gry w szachy i nie umożliwili jej dostępu do jej elementów wewnętrznych podczas ewolucji, to funkcja użytkowa mogłaby ewoluować do postaci „maksymalizacji wartości tego licznika", gdzie licznik byłby podłączony do jakiegoś czujnika w korze mierzącego liczbę wygranych gier. Jeśli następnie damy takiemu systemowi dostęp do jego wewnętrznych elementów, to słusznie zauważy, że może on znacznie lepiej zmaksymalizować swoją użyteczność, bezpośrednio zwiększając licznik, zamiast zawracać sobie głowę szachownicą. Zatem zdolność do samodzielnej modyfikacji musi być połączona z kombinacją samowiedzy i reprezentacją prawdziwych celów, a nie jakimś sygnałem zastępczym, w przeciwnym razie system jest podatny na manipulowanie sygnałem.

Nie jest jeszcze jasne, jakie mechanizmy ochronne najprawdopodobniej zastosuje SI w celu ochrony swoich systemów pomiaru użyteczności. Oczywiste jest, że zaawansowane architektury SI będą musiały radzić sobie z różnymi wewnętrznymi napięciami. SI będzie chciała mieć możliwość samodzielnej modyfikacji, jednocześnie jednak uniemożliwiając modyfikację swoich funkcji i systemów pomiarowych. Będzie chciała, aby jej podskładniki próbowały zmaksymalizować użyteczność, ale nie zrobi tego przez podrabianie lub skracanie systemów pomiarowych. Będzie chciała subkomponentów, które zbadają różne strategie, ale będą też chciały działać jako spójna, harmonijna całość. Będzie potrzebować wewnętrznych „sił policyjnych" lub „układów odpornościowych", upewniając się jednocześnie, że one same nie ulegną zepsuciu. Głębsze zrozumienie tych kwestii może również rzucić światło na strukturę ludzkiej psychiki.

SI będzie się bronić

Omówiliśmy presję na SI związaną z ochroną własnych funkcji użytkowych przed zmianami. Podobny argument pokazuje, że SI będzie uporczywie dążyć do samozachowawczości, jeśli nie zostanie wyraźnie skonstruowana inaczej. W przypadku większości funkcji użytkowych nie będzie ona działać, jeśli system zostanie wyłączony lub zniszczony. Jeśli robot szachowy zostanie zniszczony, nigdy więcej nie zagra w szachy. Użyteczność w takim przypadku będzie bardzo niska, z tego powodu systemy prawdopodobnie zrobią wszystko, by do tego nie dopuścić. Tak więc można zbudować robota szachowego z myślą o tym, że można go po prostu wyłączyć, jeśli coś pójdzie nie tak. Można się jednak zdziwić w momencie, kiedy robot zacznie zdecydowanie przeciwstawiać się takim próbom. Można spróbować zaprojektować funkcję użytkową z wbudowanymi limitami czasowymi, aczkolwiek jeśli nie zostanie to zrobione bardzo ostrożnie, system będzie zmotywowany do tworzenia systemów pomocniczych lub wykorzystywania zewnętrznych agentów, którzy nie mają limitów czasowych. Istnieje wiele strategii, które systemy mogą wykorzystać do własnej ochrony. Replikując się, system może zapewnić, że śmierć jednego z klonów nie zniszczy całkowicie systemu. Przenosząc kopie do odległych lokalizacji, może zmniejszyć podatność na lokalne, katastroficzne zdarzenia.

Istnieje wiele skomplikowanych teoretycznych zagadnień gier w rozumieniu samoobrony podczas interakcji z innymi agentami. Jeśli system jest silniejszy od innych agentów, może odczuwać presję do wykonania ataku „pierwszego uderzenia", aby zapobiegawczo zabezpieczyć się przed późniejszymi atakami ze strony innych. W przypadku gdy system jest słabszy od innych agentów, może dążyć do utworzenia infrastruktury społecznej, która będzie chronić słabych przed silnymi. Budując takie systemy, musimy bardzo uważać na tworzenie systemów, które mogą być zbyt potężne w porównaniu ze wszystkimi innymi systemami. W swojej historii ludzkość wielokrotnie doświadczała niszczącej natury władzy. Zbyt często dochodziło do przerażających aktów ludobójstwa w momencie zdobycia zbyt dużej władzy przez pojedyncze grupy.

SI będzie chciała pozyskiwać zasoby i efektywnie je wykorzystywać

Wszelkie obliczenia i działania fizyczne wymagają fizycznych zasobów przestrzeni, czasu, materii i darmowej energii. Prawie każdy cel można osiągnąć bardziej efektywnie, mając więcej tych zasobów. Maksymalizując oczekiwane użyteczności, systemy będą zatem odczuwać presję, aby zdobyć więcej tych zasobów i wykorzystać je tak efektywnie, jak to możliwe. Zasoby można uzyskać w pozytywny sposób, takie jak eksploracja, odkrywanie i handel lub za pomocą środków negatywnych, takich jak kradzież, morderstwo, przymus i oszustwo.

Niestety presja na pozyskiwanie zasobów nie uwzględnia negatywnych efektów zewnętrznych wywartych na innych. Bez wyraźnie przeciwnych celów, jest bardzo prawdopodobne, że podczas poszukiwania zasobów SI może zachowywać się jak ludzcy socjopaci. Społeczeństwa ludzkie stworzyły systemy prawne, których zadaniem jest egzekwowanie prawa własności i praw człowieka. Struktury te kierują wszelkie wysiłki związane z pozyskiwaniem w pozytywnych kierunkach, aczkolwiek muszą być stale monitorowane pod kątem ciągłej skuteczności.

Z drugiej strony wydaje się, że dążenie do efektywnego wykorzystania zasobów ma przede wszystkim pozytywne konsekwencje. Systemy będą optymalizować swoje algorytmy, kompresować dane oraz będą pracować nad wydajniejszym uczeniem się na podstawie własnych doświadczeń. Będą pracować nad optymalizacją swoich struktur fizycznych i wykonają minimalną ilość niezbędnej pracy do osiągnięcia własnych celów. Można oczekiwać, że ich fizyczne formy przyjmą eleganckie, dobrze dopasowane kształty, które tak często są tworzone w naturze.

Wnioski

Wykazaliśmy, że wszystkie zaawansowane systemy SI mogą mieć wiele podstawowych pobudek. Niezbędne jest zrozumienie tych pobudek, aby zbudować technologię, która umożliwi ludzkości budowanie pozytywnej przyszłość. Yudkowsky [13] wezwał do stworzenia „przyjaznej SI". Żeby to uczynić konieczne jest rozwinięcie nauki leżącej u podstaw „inżynierii użyteczności", która pozwoli zaprojektować funkcje użyteczności spełniające pożądane przez nas konsekwencje. Oprócz konstrukcji samych inteligentnych agentów konieczne jest również zaprojektowanie kontekstu społecznego, w którym będą funkcjonować. Struktury społeczne, które spowodują, że jednostki będą ponosiły koszty negatywnych efektów zewnętrznych, przeszłyby długą drogę w kierunku zapewnienia stabilnej i pozytywnej przyszłości. Uważam, że powinniśmy rozpocząć projektowanie „uniwersalnej konstytucji", która określałaby najbardziej podstawowe prawa przeznaczone dla jednostek oraz społeczne mechanizmy ich zapewniania w obecności inteligentnych podmiotów o zróżnicowanej strukturze. Proces ten prawdopodobnie będzie wymagał wielu iteracji, podczas których zostaną określone najważniejsze dla nas wartości oraz podejścia, które są technicznie wykonalne. Szybkie tempo postępu

technologicznego sugeruje, że kwestie te mogą wkrótce zyskać krytyczne znaczenie [14]. Udajmy się zatem w kierunku głębszego zrozumienia!

Podziękowania

Przedstawione pomysły zostały omówione z wieloma osobami, które przekazały mi cenne informacje zwrotne. Chciałbym szczególnie podziękować: Benowi Goertzelowi, Bradowi Cottelowi, Bradowi Templetonowi, Carlowi Shulmanowi, Chrisowi Petersonowi, Donowi Kimberowi, Eliezerowi Yudkowsky'emu, Ericowi Drexlerowi, Forrestowi Bennettowi, Joshowi Hallowi, Kelly Lenton, Nilsowi Nilssonowi, Rosie Wang, Shane Legg, Stevenowi Ganzowi, Susie Herrick, Tylerowi Emersonowi, Willowi Wiserowi i Zannowi Gillowi.

LITERATURA

[1] S. M. Omohundro, „The nature of self-improving artifcial intelligence." http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artifcial-intelligence/, October 2007.

[2] S. Russell and P. Norvig, Artifcial Intelligence, A Modern Approach. Pearson Education, Inc., Upper Saddle River, New Jersey, second ed., 2003.

[3] I. Marketdata Enterprises, „Self-improvement products and services," Tech: Rep., 2006.

[4] A. Mas-Colell, M. D. Whinston, and J. R. Green, Microeconomic Theory. Oxford University Press, New York, New York, 1995.

[5] J. G. Miller, Living Systems. McGraw Hill, New York, New York, 1978.

[6] L. Keller, ed., Levels of Selection in Evolution. Princeton University Press, Princeton, New Jersey, 1999.

[7] R. Trivers, Social Evolution. Benjamin/Cummings Publishing Company, Inc., San Francisco, California, 1985.

[8] R. C. Schwartz, Internal Family Systems Therapy. The Guilford Press, New York, New York, 1995.

[9] C. F. Camerer, G. Loewenstein, and M. Rabin, eds., Advances in Behavioral Economics. Princeton University Press, Princeton, New Jersey, 2004.

[10] D. Kahneman and A. Tversky, Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, Cambridge, England, 1982.

[11] S. D. Levitt and S. J. Dubner, Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. William Morrow, New York, New York, revised and expanded ed., 2006.

[12] D. Lenat, „Theory formation by heuristic search," Machine Learning, vol. 21, 1983, pp. 31 – 59.

[13] E. S. Yudkowsky, „Levels of organization in general intelligence," in Artifcial General Intelligence (B. Goertzel and C. Pennachin, eds.), Springer-Verlag, New York, 2005, pp. 389 – 501.

[14] R. Kurzweil, The Singularity is Near: When Humans Transcend Biology. Viking Penguin, New York, New York, 2005.