6 MDL destylacja inteligencji: *Poznawanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów*[*]

K. Eric Drexler

SPIS TREŚCI

Przegląd 94
Przejściowe bezpieczeństwo SI: odniesienie do trudnych przypadków 94
Porównanie ścieżek SI wysokiego i niskiego ryzyka 96
Wiedza, nauka i destylacja MDL 96
Dlaczego system SI dąży do MDL zamiast do inteligencji? 97
Pominięcie treści językowych, pominięcie wiedzy o dziedzinie 97
Pominięcie planów zorientowanych na zewnątrz 99
Destylacja pasuje do obecnej praktyki badawczej 100
Od destylacji MDL po narzędzia SI z obsługą superinteligencji 100
Specjalizacja i skład 101
Sposoby i wyzwania związane z wdrażaniem specjalizacji 102
Modularne architektury specjalistyczne 102
Perspektywy i kierunki badań 103
Niektóre otwarte pytania 104
Streszczenie 106
Podziękowania 107
Aneks: Bezpieczne architektury dla superinteligentnej inżynierii 107

[*] Przedruk z Drexler, K.E. 2015. „MDL Intelligence Distillation: Exploring strategies for safe access to superintelligent problem-solving capabilities", *Technical Report* #2015-3, Future of Humanity Institute, Oxford University: s. 1–17. Przedruk za zgodą autora.

PRZEGLĄD

Technologie SI mogą osiągnąć próg szybkiej, otwartej, rekurencyjnej poprawy, zanim będziemy przygotowani na wyzwania związane z pojawieniem się superinteligentnych agentów SI.* Jeśli taka sytuacja nastąpi, może okazać się niezwykle ważne zastosowanie metod zmniejszania ryzyka sztucznej inteligencji, dopóki bardziej kompleksowe rozwiązania nie zostaną zrozumiane i gotowe do wdrożenia. Jeśli metody redukcji ryzyka mogą przyczynić się do tych kompleksowych rozwiązań, tym lepiej.

Podstawowa technika zmniejszania ryzyka sztucznej inteligencji obejmowałaby możliwości rekurencyjnego doskonalenia sztucznej inteligencji do określonego zadania. Jest to proces nazwany „destylacją inteligencji", w którym miarą inteligencji SI jest minimalizacja długości opisu implementacji, które same są zdolne do otwartej poprawy rekurencyjnej.

Oddzielając wiedzę od zdolności uczenia się, destylacja inteligencji może wspierać strategie wdrażania wyspecjalizowanych, mało ryzykownych, a jednocześnie superinteligentnych mechanizmów rozwiązywania problemów: destylacja może ograniczać początkową ilość informacji, pomiar wiedzy może ograniczać wprowadzanie informacji podczas uczenia się, protokoły punktu kontrolnego/restartu mogą ograniczać przechowywanie informacji dostarczanych w połączeniu z zadaniami. Opierając się na tych metodach i ich produktach funkcjonalnych, zestawy mechanizmów z superinteligentnymi kompetencjami dziedzinowymi do rozwiązywania problemów mogą zostać potencjalnie połączone w celu wdrożenia wysoce wydajnych systemów, które nie mają cech charakterystycznych dla silnej i ryzykownej SI. W aneksie opisano, w jaki sposób można zastosować tę strategię do wdrożenia superinteligentnych, interaktywnych systemów inżynieryjnych przy minimalnym ryzyku.

Strategie destylacji/specjalizacji/składu implikują szerokie pytania dotyczące potencjalnego zakresu bezpiecznych zastosowań zdolności SI opartej na superinteligencji. Ponieważ strategie umożliwiające destylację mogą oferować praktyczne środki zmniejszania ryzyka SI przy realizacji ambitnych zastosowań, dalsze badania w tym obszarze mogłyby wzmocnić powiązania między społecznościami zajmującymi się opracowywaniem SI i badaniami nad bezpieczeństwem SI.

PRZEJŚCIOWE BEZPIECZEŃSTWO SI: ODNIESIENIE DO TRUDNYCH PRZYPADKÓW

W książce *Superintelligence* (Oxford University Press, 2014) Nick Bostrom analizował szereg głębokich problemów związanych z potencjalnym pojawieniem się superinteligentnych

* Ostatnia książka Nicka Bostroma *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), zapewnia najszerszą i najgłębszą analizę tych wyzwań. Niniejszy dokument przeznaczony jest dla publiczności, która ma ogólną wiedzę na temat rozważań i problemów poruszonych w *Superintelligence.Disclaimer*. Po rozmowie na ten temat w Future of Humanity Institute 4 grudnia 2014 roku Anders Sandberg zasugerował, żebym napisał krótkie streszczenie, ale pomimo że dokument ten jest zgodny z treścią rozmowy, to zawiera skrócone opisy pojęć i nie zawiera aparatu cytowania naukowego. Uwaga historyczna: Moje obawy dotyczące ryzyka SI, które koncentrują się na wyzwaniach związanych z długoterminowym zarządzaniem SI, rozpoczęły się wraz z początkiem moich badań zaawansowanych technologii molekularnych, około 1977 roku. Przypominam sobie późniejszą rozmowę z Marvinem Minskym, ówczesnym przewodniczącym mojej komisji doktorskiej (około 1990 roku), która pogłębiła moje zrozumienie niektórych kluczowych kwestii. W odniesieniu do hierarchii celów Marvin zauważył, że głównym zadaniem nauczenia się języka przez niemowlaka jest podcel napicia się wody, a zamiana zasobów wszechświata w komputery to potencjalny podcel maszyny próbującej doskonale grać w szachy. Przedstawione tutaj pomysły pojawiły się jako podcele proponowanych strategii zarządzania niewiarygodnymi systemami sztucznej inteligencji, które przedstawiłem Marvinowi w tym samym czasie. Zasugerował, żebym to opisał. Na razie zwlekam.

jednostek SI i sugeruje, że odpowiednie rozwiązania mogą być znacznie opóźnione. Jeśli technologie SI osiągną próg szybkiej, otwartej, rekurencyjnej poprawy, zanim będziemy w stanie w pełni rozwiązać problemy omówione w *Superintelligence*, to tymczasowe strategie kształtowania i zarządzania powstającą superinteligencją mogą być kluczowe.

Za referencyjny problem/sytuację przyjęto następujące warunki:

1. Technologia SI osiągnęła próg szybkiej, otwartej, rekurencyjnej poprawy.
2. Treść i mechanizmy powstających superinteligentnych systemów są skutecznie nieprzejrzyste.
3. Ciągłe naciski na zastosowania SI zapewniają szerokie wykorzystanie superinteligencji.
4. Żadne w pełni adekwatne rozwiązanie problemów stwarzanych przez superinteligentne jednostki nie jest gotowe do wdrożenia.

Warunki od 1 do 4 są trudne, ale zgodne z potencjalnie potężnymi i dostępnymi strategiami redukcji ryzyka. Te strategie można oczywiście zastosować w mniej wymagających okolicznościach.

Rozważając siłę punktu 3, należy wziąć pod uwagę ciągłą presję na stosowanie zaawansowanych zdolności SI, w tym samą dynamikę konkurencyjnych badań i rozwoju. Zastosowania superinteligencji mogą być nie tylko wyjątkowo zyskowne, ale mogą znacznie zwiększyć wiedzę naukową, globalne bogactwo materialne, zdrowie ludzkie, a może nawet prawdziwe bezpieczeństwo. Ponieważ nierozsądne byłoby zakładanie, że pojawiająca się superinteligencja nie będzie stosowana, istnieje dobry powód, aby szukać środków do wdrażania zastosowań o niskim ryzyku.

Z perspektywy redukcji ryzyka przejściowe środki bezpieczeństwa SI oferują kilka potencjalnych korzyści:

1. Mogą wydłużyć czas przeznaczony na badanie podstawowych problemów związanych z długoterminową kontrolą SI.
2. Mogą umożliwić eksperymentowanie z działającymi i potencjalnie zaskakującymi technologiami SI.
3. I być może najważniejsze, mogą umożliwiać zastosowanie superinteligentnych mechanizmów rozwiązywania problemów do kwestii zarządzania superinteligencją.

TABELA 6.1.
Potencjalne ścieżki do niebezpiecznych agentów SI versus narzędzia SI niskiego ryzyka

Potencjalna ścieżka do niebezpiecznych agentów SI	Potencjalna ścieżka do narzędzi SI niskiego ryzyka
Otwarta, niekierowana, rekurencyjna poprawa skutkuje pojawieniem się superinteligentnego systemu. Superinteligencja zdobywa szeroką światową wiedzę, opracowuje wyraźne, dalekosiężne cele, opracowuje plany działania o zasięgu globalnym, stosuje skuteczne środki do realizacji swoich planów.	Zmierzone, powtarzalne, rekurencyjne doskonalenie skutkuje pojawieniem się superinteligentnych uczniów o minimalnej zawartości, którzy umożliwiają wykształcenie systemów dysponujących specjalistyczną wiedzą. Systemy te badają rozwiązania zadanych problemów, wykonują obliczenia przy użyciu przydzielonych zasobów, wykonują przydzielone zadania, udzielając odpowiedzi.

Porównanie ścieżek SI wysokiego i niskiego ryzyka

W tabeli 6.1 zestawiono potencjalną ścieżkę rozwoju SI prowadzącą do powstania agenta SI wysokiego ryzyka, z proponowaną ścieżeką rozwoju i stosowania superinteligentnych możliwości za pomocą środków, które potencjalnie mogłyby wyeliminować to ryzyko.

Należy zauważyć, że zasadniczym aspektem części 1 ścieżki niskiego ryzyka jest standardowa praktyka badawcza: przechowywanie kopii zapasowych lub punktów kontrolnych stanu systemu podczas programowania oraz rejestrowanie kroków prowadzących do kolejnego interesującego wyniku. Wspólnie praktyki te umożliwiają śledzenie i modyfikację ścieżek rozwoju podczas badania charakterystyk stanów pośrednich.

W poniższej dyskusji założono, że wzdłuż ścieżek zmierzających w kierunku potencjalnie ryzykownej superinteligencji zdolność do rekurencyjnego doskonalenia poprzedza agent SI o wysokim ryzyku lub przynajmniej, że warunek ten można ustalić przez kontrolowaną przebudowę możliwości rekurencyjnych ulepszeń wzdłuż alternatywnych ścieżek, zaczynając od wczesnego i bezproblemowego punktu kontrolnego. Ten warunek zapewnia, że strategie kontrolne mogą być stosowane w kontekście innym niż przeciwny (rysunek 6.1).

RYSUNEK 6.1. Schemat działania destylacji MDL mającej na celu wytworzenie i następnie rozwinięcie zwartych systemów uczenia się ogólnego zastosowania

WIEDZA, NAUKA I DESTYLACJA MDL

Na ścieżce niskiego ryzyka przedstawionej w tabeli 6.1 kluczowy jest krok 2. Polega na stworzeniu szczególnego rodzaju superinteligencji, superinteligentnego ucznia o minimalnej ilości informacji. W jaki sposób można to osiągnąć?

Z założenia referencyjna, problematyczna sytuacja zawiera systemy SI zdolne do wdrażania systemów SI bardziej inteligentnych od nich samych.

Odpowiednio sprawny bazowy system SI może następnie zostać podany jako argument operatorowi udoskonalania SI, który stosuje bazową SI do przepisania drugiego systemu SI w celu stworzenia trzeciego, bardziej inteligentnego systemu SI: