MDL destylacja inteligencji:

# Poznawanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów

K. Eric Drexler

## 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 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

**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. |

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.

### *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 ścieżką 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).

## 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?

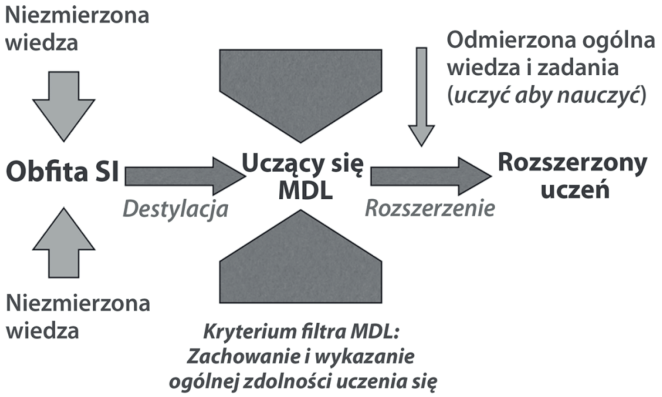
Z założenia referencyjna, problematyczna sytuacja zawiera systemy SI zdolne do wdrażania systemów SI bardziej inteligentnych od nich samych.

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:

1. Ulepsz (*bazowa SI, docelowa SI*, wskaźnik (*zadania, inteligentniejsza*)) → *inteligentniejsza SI*,

gdzie „inteligentniejszy" jest zdefiniowany w kategoriach odpowiednio ogólnych wskaźników wydajności zadania. Prawdopodobnie możemy sparametryzować ten operator za pomocą dowolnego z szeregu wskaźników poprawy, w tym wskaźników dotyczących ilości informacji produktu:

2. Ulepsz (*bazowa SI, docelowa SI*, wskaźnik (*zadania*, *mniejsza*)) → *mniejsza SI*.




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

W tym przypadku poprawa polega na zmniejszeniu rozmiaru wynikowej SI pod warunkiem zapewnienia odpowiedniej wydajności wykonywanych zadań.

Zadania kryterialne mogą wymagać, aby wynikowa SI spełniała szeroki zakres testów wydajności *po procesie uczenia się z odpowiednich programów nauczania*. Biorąc pod uwagę wystarczająco ogólną, superinteligentną docelową SI, odpowiednio dobrany zestaw zadań kryterialnych może zapewnić, że wynikowy system SI będzie ogólnym, superinteligentnym uczniem.

W referencyjnej problematycznej sytuacji, gdzie zakłada się nieprzejrzystą, mocno ulepszającą się technologię SI, można zastosować operator poprawy w następujący sposób:

3. Popraw (*początkowa SI, początkowa SI*, wskaźnik (*zadania, min-MDL)) → MDL-wydestylowana-SI*, gdzie wynikowa „MDL-wydestylowana-SI" ma dwie kluczowe właściwości:

- Kryteria wykonania zadania zapewniają, że podobnie jak początkowa SI, produkt może się uczyć bez ograniczeń i rekurencyjnie się doskonalić.
- Wskaźnik MDL zapewnia, że w ramach ograniczeń zasobów produkt jest najbardziej kompaktowym takim systemem, jaki może zbudować początkowa SI.

### *Dlaczego system SI dąży do MDL zamiast do inteligencji?*

System SI poprawiający SI może naturalnie wykonywać szereg zadań związanych z implementacją kompaktowej SI, opracowując kompaktowe wersje systemów MDL, które mogą nauczyć się grać w szachy lub nauczyć się pokonywać Watsona, grając w Jeopardy! itd. Opracowywanie kompaktowych wersji systemów zdolnych do otwartego uczenia się i rekurencyjnego doskonalenia jest zasadniczo podobnym rodzajem zadania wdrożeniowego: optymalizacją systemu pod kątem zwartości zgodnie z ogólnymi kryteriami uczenia się i wydajności. Należy pamiętać, że zadania tego rodzaju nie są związane z refleksyjnymi obawami związanymi z samodoskonaleniem.

W zakresie, w jakim mogą pojawić się obawy dotyczące problematycznego zachowania strategicznego w nieprzezroczystych, źle scharakteryzowanych systemach SI, obawy te można potencjalnie rozwiązać przez na przykład ponowne uruchomienie procesu poprawy SI z bezproblematycznego punktu kontrolnego i interpolację kolejnych etapów destylacji MDL.

### *Pominięcie treści językowych, pominięcie wiedzy o dziedzinie*

„Pomiar wiedzy", kontrolowanie wprowadzania informacji, oferuje potężną technikę ograniczania zawartości destylowanych systemów MDL. Zastanówmy się nad językiem:

Niemowlęta pokazują, że inteligentne systemy mogą osiągnąć ogólne możliwości uczenia się bez uciekania się do początkowego wyposażenia w treści językowe (tzn. bez znajomości konkretnej gramatyki lub słownictwa). W szczególności ogólna umiejętność uczenia się języka jest konsekwencją silnych priorytetów dotyczących abstrakcyjnej struktury języka w połączeniu z bardzo słabymi priorytetami dotyczącymi konkretnych treści językowych.

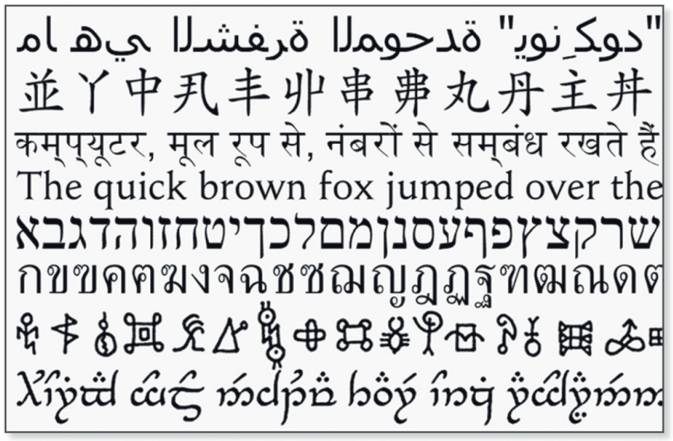
Destylacja uczniów MDL w naturalny sposób pomija słownictwo, ponieważ jest ono nieporęczne i łatwo się go nauczyć lub je zainstalować. Należy zauważyć, że słownictwa nie można odgadnąć bez konkretnej wiedzy. Jakiekolwiek domysły nie pozwoliłyby na uzyskanie słownika chińskiego, angielskiego, klingońskiego lub chicomuceltec? Słownictwa, podobnie jak innych historycznie zależnych informacji językowych (np. rysunek 6.2), nie można wywnioskować ze źródeł niezależnych od języka.

Podobne uwagi dotyczą historycznie zależnych zasobów wiedzy, które stanowią większość treści przeważającej części dziedzin akademickich (np. nauk biologicznych), oraz wiedzy (np. chemii) zależnej od parametrów fizycznych, takich jak masa elektronu. Rodzaje wiedzy, które koniecznie (choć być może domyślnie) zostaną zatrzymane przez ucznia MDL, prawdopodobnie mieszczą się w zakresie dyscyplin akademickich zwanych „naukami formalnymi" z tabeli 6.2. Rozwój profesorów od stadium niemowlęcia pokazuje, że nieprzewidziane wyroki i ogólne mechanizmy zapewniają wystarczającą podstawę do otwartego uczenia się.

Biorąc pod uwagę, że pominięto zbiór informacji warunkowych, odpowiednie ograniczenia dotyczące wprowadzania informacji mogą uniemożliwić ich późniejsze pozyskanie. Ocena ograniczeń wynikających z konkretnej polityki pomiaru wiedzy będzie jednak wymagała uwzględnienia nie tylko wiedzy bezpośredniej, ale także wywnioskowanej. Ograniczenia wnioskowania będą czasem jasne, ale na przykład przy analizie próbek nieformalnej wiedzy o świecie zakres wiedzy wnioskowanej może być niezwykle trudny do oceny.

**Tab. 6.2 Dyscypliny akademickie odnoszące się do bardzo różnych zadań**

| 1. Humanistyczne | 3. Nauki przyrodnicze | 5. Zawody |
|---|---|---|
| 1.1. Historia człowieka<br>1.2. Lingwistyka<br>1.3. Literatura<br>1.4. Sztuka<br>1.5. Filozofia<br>1.6. Religia | 3.1. Biologia<br>3.2. Chemia<br>3.3. Nauki o Ziemi<br>3.4. Fizyka<br>3.5. Nauki o kosmosie | 5.1. Rolnictwo<br>5.2. Architektura<br>5.3. Biznes<br>5.4. Teologia<br>5.5. Pedagogika<br>5.6. Inżynieria<br>5.7. Środowiskowe...<br>5.8. Rodzinne...<br>5.9. Kultura fizyczna...<br>5.10. Dziennikarstwo...<br>5.11. Prawo<br>5.12. Bibliotekarstwo<br>5.13. Medycyna<br>5.14. Nauki wojskowe<br>5.15. Administracja publiczna<br>5.16. Prace społeczne<br>5.17. Transport |
| **2. Nauki społeczne** | **4. Nauki formalne** | |
| 2.1. Antropologia<br>2.2. Archeologia<br>2.3. Studia obszarowe<br>2.4. Kulturalne...<br>2.5. Ekonomia<br>2.6. Studia gender<br>2.7. Geografia<br>2.8. Nauki polityczne<br>2.9. Psychologia<br>2.10. Socjologia | 4.1. Matematyka<br>4.2. Informatyka<br>4.3. Logika<br>4.4. Statystyka<br>4.5. Nauki systemowe | |



Rys. 6.2. Warunkowe informacje językowe

### *Pominięcie planów zorientowanych na zewnątrz*

Reprezentowanie planów wymaga informacji, a w zakresie, w jakim plany nie są istotne dla zadania, destylacja będzie dążyć do usunięcia informacji, które je zawierają. W szczególności plany, które są zarówno specyficzne, jak i zorientowane na świat zewnętrzny muszą zawierać istotne informacje warunkowe, które jak już pokazano, nie są potrzebne dla ogólnych możliwości uczenia się.

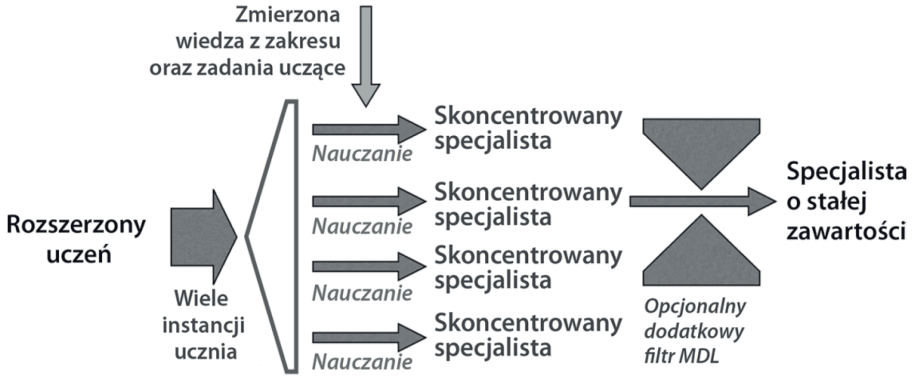
Można sprzeciwić się temu, że w możliwej do uniknięcia sytuacji kontradyktoryjnej problematyczne plany mogą zostać osadzone w strukturach istotnych dla zadania w sposób, który z założenia utrudnia ich identyfikację i usunięcie. Jednak proces destylacji z obsługą superinteligencji prawdopodobnie byłby w stanie zastosować świeże, zwarte struktury o podobnej funkcjonalności. Niepotrzebnie złożone struktury nie muszą zostać rozumiane, żeby zostały odrzucone.

### *Destylacja pasuje do obecnej praktyki badawczej*

Destylacja MDL ma na celu oddzielenie wiedzy od możliwości uczenia się, a w dzisiejszym uczeniu maszynowym ta separacja już się utrzymuje. Systemy głębokiego uczenia mogą mieć zaskakująco kompaktowe abstrakcyjne specyfikacje i ciągle mogą być uczone przy użyciu gigabajtów danych w celu wytworzenia systemów z megabajtami nieprzejrzystej, numerycznej treści.

W uczeniu maszynowym oddzielanie wiedzy od umiejętności uczenia się jest zarówno dobrą nauką, jak i dobrą inżynierią:

- Oddzielenie treści wiedzy od umiejętności uczenia się ułatwia ludzkie zrozumienie procesów uczenia się i ich produktów.
- Szkolenie systemów uczących bez zawartości ze znanymi zestawami danych umożliwia powtarzalność i analizę porównawczą podczas projektowania.




Rys. 6.3. Ogólne podejście do produkcji specjalistycznych systemów z MDL destylowanych (a następnie rozszerzonych) systemów uczących (rysunek 6.2)

- Szkolenie systemów uczących bez zawartości minimalizuje obciążenia zależne od ścieżki i umożliwia różnorodne zastosowania określonych metod uczenia się.
- Zasady MDL często poprawiają uogólnianie od przykładów uczenia się do danych wykorzystywanych następnie podczas testowania, sprawdzania i zastosowania.

Na progu rekurencyjnego ulepszania SI destylację MDL można zastosować do oddzielenia wiedzy od zdolności uczenia się, nawet jeśli są ze sobą powiązane, i tym samym można zapewnić sposób zachowania lub odzyskania korzyści naukowych, inżynieryjnych i bezpieczeństwa obecnych praktyk badawczych.

## OD DESTYLACJI MDL PO NARZĘDZIA SI Z OBSŁUGĄ SUPERINTELIGENCJI

Wdrożenie trzeciego kroku wzdłuż proponowanej ścieżki niskiego ryzyka prowadzącej do narzędzi SI (tabela 6.1) wymaga nauczenia superinteligentnych uczniów o minimalnej zawartości za pomocą ogólnej („uczyć, aby nauczyć”), a następnie specjalistycznej wiedzy w celu opracowania specjalistycznych systemów SI przeznaczonych dla konkretnych zastosowań. Na rysunku 6.3 przedstawiono ogólne podejście.

Nauczanie destylowanego, efektywnie pustego ucznia MDL umożliwia pomiar i audyt początkowej ilość wiedzy na temat wynikowych produktów SI. Podejście to ogranicza część 2 referencyjnej problematycznej sytuacji, potencjalną nieprzejrzystość zawartości wiedzy wynikowych superinteligentnych systemów. Destylacja i pomiar wiedzy mogą ograniczać ilość wiedzy bez względu na jej reprezentację.

Specjalistyczne kompetencje mogą być wąskie i jednocześnie potężne. Jako przykłady można wymienić superinteligentne maszyny do dowodzenia twierdzeń, architektów komputerowych oraz systemy o superinteligentnych kompetencjach inżynieryjnych do rozwiązywania wspólnych problemów strukturalnych, mechanicznych, termicznych i aerodynamicznych przy projektowaniu samolotów hipersonicznych. Nauczanie ściśle skoncentrowanych specjalistów, pomijając bezpośrednią lub ukrytą znajomość języka, polityki i geofizyki, może wymagać uwagi, ale nie zawsze musi być trudne.

W niektórych dziedzinach zadania będą zawierać potencjalnie znaczącą informację o pozornie niepowiązanych aspektach świata zewnętrznego. Informacje przekazywane przez strumień zadań nie muszą się jednak kumulować, ponieważ systemy rozwiązywania problemów nie muszą przekazywać informacji z poprzednich instancji (np. punktów kontrolnych). Bardziej swobodna polityka umożliwiłaby kumulatywne uczenie się w postaci kanonicznych reprezentacji rezultatów zadań, takich jak twierdzenia matematyczne, obwody cyfrowe lub nowe konfiguracje mechaniczne, innymi słowy, zwięzłe reprezentacje wiedzy związanej z zadaniami.

### *Specjalizacja i skład*

Wysoko wyspecjalizowane jednostki zazwyczaj zajmują się tylko pewnymi częściami problemów, co gwałtownie ogranicza ich zastosowania w izolacji. Naturalne jest zatem łączenie specjalistów z wąskich dziedzin w celu budowania systemów modułowych, które, pomimo że nadal są wyspecjalizowane, mają szersze zastosowanie.

Istnieją rozległe precedensy dotyczące budowania mechanizmów szerokiego rozwiązywania problemów na podstawie wyspecjalizowanych elementów, na przykład:

- Układy nerwowe łączące korę wzrokową, słuchową i ruchową.
- Zespoły inżynierów złożone z różnych ludzkich specjalistów.
- Gospodarki rynkowe z szerokim podziałem pracy i wiedzy.
- Skomplikowane architektury oprogramowania złożone z komponentów modułowych.

Jak sugerują te przykłady, systemy złożone z różnych specjalistów mogą mieć niezwykle szerokie możliwości zastosowania. Jednak w kontekście bezpieczeństwa SI potencjał ten uwypukla możliwość powstania systemów o znacznym ryzyku, skomponowanych z bezpiecznych komponentów. Tak więc, pomimo że systematyczną eksplorację potencjalnych systemów opartych na superinteligencji można rozpocząć od zbadania środków wdrażania wyspecjalizowanych komponentów, należy zwrócić uwagę na pytania dotyczące wyłaniających się właściwości, bezpieczeństwa i ryzyka nie tylko w wielu dziedzinach i zadaniach, ale w kontekście alternatywnych architektur modułowych.

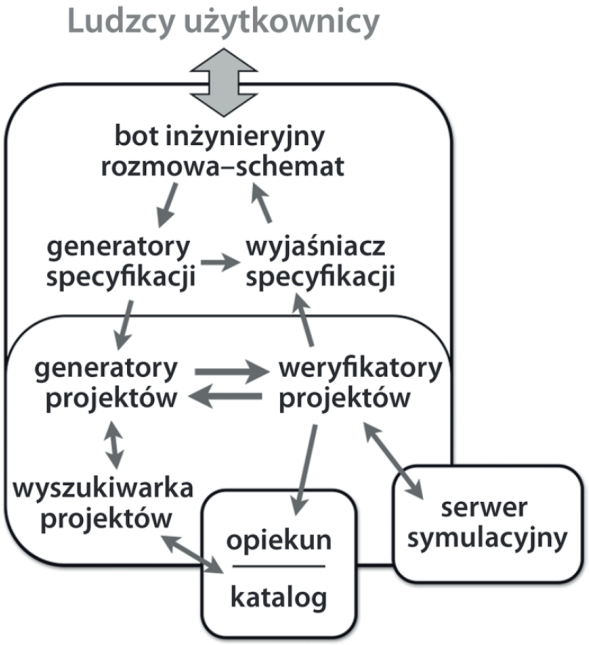
### *Sposoby i wyzwania związane z wdrażaniem specjalizacji*

W niektórych obszarach pomiar wiedzy może ustanowić wyraźne ograniczenia w potencjalnej wiedzy interferencyjnej między dziedzinami. W innych obszarach potencjalne wnioskowanie pomiędzy dziedzinami może być szerokie i nieprzewidywalne. Biorąc pod uwagę zakres potencjalnego wnioskowania, należy jednak zauważyć, że możliwości uczenia

się specjalistycznego systemu można ograniczyć przez destylację wtórną (po nauczaniu).Uzyskując systemy nieuczące oraz, jak wspomniano wcześniej, takie, których kumulatywne uczenie się związane z zadaniami może być bezpośrednio ograniczone przez zasady punktu kontrolnego/restartu.

Oprócz ograniczeń wiedzy *jako takiej*, specjalistyczną SI można dalej kształtować i ograniczać za pomocą wskaźników destylacji, które optymalizują kompromisy między zasobami i wydajnością w odniesieniu do strumieni zadań specyficznych dla dziedziny, ograniczając w ten sposób zakres innych funkcji. Dodatkowo może to być wykonane także przez stałe interfejsy, które wprowadzają opisy zadań i wyniki do reprezentacji specyficznych dla dziedziny (np. wyrażenia matematyczne, specyfikacje inżynierii fizycznej), w istocie interfejsy API.

Podobnie jak w przypadku wnioskowania stosowanego do zasobów wiedzy, czasami trudno będzie ocenić zakres kształtowania i specjalizacji, które mogą być wywołane przez optymalizację wydajności zadania, kontrolę strumienia zadań i interfejsy API specyficzne dla dziedziny. Techniki te poszerzają bogaty zestaw narzędzi, które prowadzą do szerokiego zakresu pytań dotyczących specjalizacji, bezpieczeństwa i ryzyka w kontekście konkretnych dziedzin, zadań i architektur systemów obsługiwanych przez specjalistów.

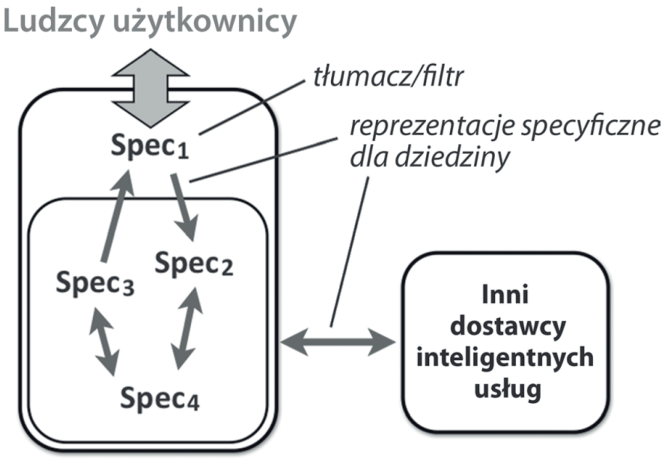



Rys. 6.5. Destylowani specjaliści przygotowani do wdrożenia systemu o szerokich kompetencjach inżynieryjnych

### *Modularne architektury specjalistyczne*

Na rysunku 6.2 przedstawiono ogólny schemat komponowania destylowanych specjalistów w celu wdrożenia systemów o bardziej ogólnych możliwościach (rysunek 6.4).

Należy pamiętać, że zadanie polegające na komunikacji pośredniczącej pomiędzy użytkownikami ludzkimi a specjalistami z różnych dziedzin może być realizowane przez specjalny interfejs komunikacyjny, który umożliwia użytkownikom przesyłanie i wyjaśnianie opisów zadań w drodze dyskusji w podzbiorze języka naturalnego dla danej dziedziny (potencjalnie wspomagany grafiką interaktywną), jednocześnie wymieniając reprezentacje dotyczące konkretnych zadań z systemem złożonym ze specjalistów z danej dziedziny. Specjalizacja polega również na rozbijaniu zadań na węższe podzadania, podobnie jak tłumaczenie wyników na formy zrozumiałe dla ludzkiego użytkownika.




Rys. 6.4. Schematyczna architektura systemu powiązanych specjalistów przetłumaczona i odfiltrowana przez specjalistę poprzez interfejs użytkownika. Patrz także rysunek 6.5

Każdy lub wszyscy tacy specjaliści mogą zostać niezdolni do długoterminowego, kumulatywnego uczenia się poprzez zainicjowanie każdego zadania za pomocą systemu w ustalonym stanie początkowym. Byłoby to całkiem naturalne: unikanie modyfikacji zawartości systemu pomiędzy zadaniami ma tę zaletę, że zapewnia spójne zachowanie, które może być zarówno dobrą praktyką inżynierską, jak i pomocą w debugowaniu.

W Aneksie przedstawiono bardziej konkretny przykład modułowej kompozycji złożonej ze specjalistów dla ważnego przypadku projektowania inżynierskiego (rysunek 6.5).

## PERSPEKTYWY I KIERUNKI BADAŃ

Destylacja inteligencji, pomiar wiedzy, specjaliści zorientowani, punkt kontrolny/restartu oraz skład modułowy to ogólne środki kontroli z wieloma potencjalnymi instancjami i wspólnymi zastosowaniami. Pojęcia te, rozpatrywane zarówno indywidualnie, jak i jako całość, nasuwają pytania dotyczące nie tylko potencjalnego zakresu, implementacji i zastosowania, ale także dotyczące skutecznych metodologii badania tego zakresu pytań pod kątem potencjalnie krytycznych decyzji prowadzących do superinteligencji.

### *Niektóre otwarte pytania*

Jak zinterpretować „minimalną długość opisu”? Ze względów praktycznych opis w odniesieniu do maszyny Turinga nie jest odpowiedni. Zamiast tego opis może być wyrażony w języku wysokiego poziomu lub specyfikacji wykonawczej i może zawierać spakowane, nieprzezroczyste algorytmy wybrane z danej biblioteki, redukując w ten sposób wiele opisów algorytmów do indeksów macierzy. Należy pamiętać, że zestaw rozważań dotyczących ograniczeń zasobów, treści programu nauczania i elastycznej koncepcji „uczyć, aby nauczyć” są łącznie istotne przy formułowaniu odpowiednich wskaźników długości opisu.

Ponadto, w jaki sposób można modelować ryzyko związane z SI i zależności od środków kontrolnych? Jeśli można zastosować różnorodne techniki w celu zmniejszenia różnych aspektów ryzyka silnych jednostek, w jaki sposób można modelować redukcję ryzyka osiągniętą za pomocą wielu technik? Jakie środki kontrolne mogą być modelowane jako probabilistyczne, niezależne i multiplikatywne? Które z nich są słabe, jeśli są stosowane osobno, a jednocześnie mocne, gdy są stosowane w połączeniu z innymi? Które dzielą typowe tryby awarii?

Jakie są w takich ramach progi niebezpiecznej agencji? Co stanowi granicę pomiędzy narzędziami SI o niskim ryzyku i wysoko ryzykownymi agentami SI? Kiedy wnioskowanie na podstawie bazy wiedzy może dać nieoczekiwaną wiedzę i ewentualnie nieoczekiwane możliwości? Jak szerokie są regiony, które można śmiało uznać za bezpieczne?

Perspektywy bezpiecznych zastosowań superinteligencji sugerują dalsze otwarte pytania:

- W jaki sposób moglibyśmy wykorzystać superinteligentne narzędzie do dowodzenia twierdzeń?
- Na jakie pytania mogą odpowiedzieć wyspecjalizowane superinteligentne systemy?
- Czy superinteligencja może pomóc nam rozwiązać problemy związane z SI?
- Czy moglibyśmy zbudować wielostronne gry wśród *niezaufanych* superinteligentnych systemów, aby uzyskać wiarygodne rozwiązania problemów silnych jednostek SI?

**Tabela 6.3 Zakres tematów technicznych i uwag**

| **Potencjalne obawy dotyczące progu SI** | **Architektury specjalistyczne** |
|---|---|
| • Monitorowanie pojawiających się możliwości<br>• Zastosowania punktu kontrolnego/restartu | • Faktoring kompetencji<br>• Modułowe wzory kompozycji |
| **Proces destylacji i wskaźniki** | **Projektowanie interfejsów informacyjnych** |
| • Zastosowania iteracyjnej destylacji<br>• Destylacja zależna od dziedziny | • Filtrowanie na ludzkich interfejsach<br>• Monitorowanie na interfejsach wewnętrznych |
| **Dziedziny i programy nauczania** | **Ryzyko specyficzne dla zastosowania** |
| • Ogólne i specjalistyczne programy nauczania<br>• Nauczanie kontra ładowanie bazy danych | • Robotyka interaktywna na całym świecie<br>• Dostęp do Internetu i interakcja |
| **Podział wiedzy** | **Ryzyko wystąpienia agencji** |
| • Dziedziny i podziały<br>• Niejednoznaczności w zakresie wiedzy | • Granice ryzyka agencji<br>• Bezpieczny skład ryzykownych agentów |

**Tabela 6.4 Zakres rozważań dotyczących badań nad SI**

| **Obecne praktyki badawcze SI** | **Oczekiwane obawy ekonomiczne** |
|---|---|
| • Ocena obecnej praktyki<br>• Destylacja jako dobra nauka<br>• Ocena bieżących zastosowań<br>• Prekursory ryzykownych jednostek SI | • Zmniejszenie niepewności ochronnych<br>• Minimalizacja kosztów samoochrony<br>• Minimalizowanie opóźnień samoochrony<br>• Włączanie bezpiecznych zastosowań |

**Tabela 6.5 Zakres rozważań dotyczących bezpieczeństwa SI**

| **Wypełnianie luki w programach badań nad SI** | **Realizacja celów długoterminowych** |
|---|---|
| • Obawy krótkoterminowe kontra długoterminowe<br>• Problemy konkretne kontra abstrakcyjne<br>• Zastosowania kontra badanie ryzyka | • Wzbogacanie wszechświata koncepcyjnego<br>• Poszukiwanie ścieżek przez przemiany<br>• Poszukiwanie czynników umożliwiających pełne rozwiązania |
| **Poszerzenie wsparcia badań ryzyka** | |
| • Angażowanie nowych badaczy<br>• Rozwiązanie szerszej gamy problemów<br>• Motywowanie szerszej gamy fundatorów | |

W tabeli 6.3 przedstawiono wybrane tematy techniczne wymagające dalszych badań. Obejmują one techniki monitorowania zdolności podczas opracowywania SI przez określone środki kontroli SI i zakres ich zastosowania.

Przechodząc do obaw innego rodzaju, w tabeli 6.4 przedstawiono szereg rozważań związanych z potencjalnymi ścieżkami rozwoju SI, w szczególności kluczowe obawy, które mogą powstać w kontekście bieżących projektów badawczych i rozwojowych, w tym potencjalne koszty, niepewności, ograniczenia i opóźnienia spowodowane wdrażaniem alternatywnych polityk ochronnych. Podane tutaj podejście do tymczasowego bezpieczeństwa SI sugeruje możliwość opracowania konkretnych i strawnych porad, które będą zgodne z istniejącą praktyką badawczą, a w szczególności metodami, które oddzielają możliwości uczenia się od nauczonej treści, oferując jednocześnie możliwość identyfikacji ścieżek niskiego ryzyka do szeregu satysfakcjonujących zastosowań superinteligentnych technologii SI.

Przechodząc do badań nad ryzykiem związanym z SI, badania nad przejściowym zarządzaniem ryzykiem SI mogą potencjalnie pomóc wypełnić lukę, nie tylko w rzeczywistych technikach kontroli ryzyka (np. opóźnienie w przygotowaniu, które określa referencyjną problematyczną sytuację, część „Przejściowe bezpieczeństwo SI: odniesienie do trudnych przypadków"), ale także pomiędzy społecznościami badawczymi zajmującymi się SI zorientowanymi na ryzyko oraz na rozwój. Społeczności te obecnie znajdują się w dobrym kontakcie, aczkolwiek mógłby on zostać jeszcze bardziej wzmocniony.

Badania ryzyka koncentrujące się na nierozwiązanych problemach dotyczących superinteligentnych jednostek SI są z natury abstrakcyjne i długoterminowe, a zatem mają niewiele praktycznych implikacji dla obaw współczesnych twórców SI. Z drugiej strony, zapytanie o przejściowe strategie bezpieczeństwa SI (tabela 6.5) koncentruje się na badaniu terytorium pomiędzy dzisiejszymi celami badawczymi a długoterminowymi obawami. Może to być źródłem porad istotnych dla problemów krótkoterminowych, a być może pomóc nam w przerobieniu i przeformułowaniu problematycznych sytuacji na potrzeby badań nad długoterminową kontrolą ryzyka SI.

## STRESZCZENIE

W znanej i trudnej referencyjnej problematycznej sytuacji technologia SI osiągnęła próg szybkiej, rekurencyjnej poprawy,

opierając się na nieprzejrzystych, słabo zrozumianych systemach SI. Jednocześnie ze względu na presję ekonomiczną i inne czynniki pojawiająca się superinteligencja została zastosowana do rozwiązywania praktycznych problemów, zanim jeszcze zagadnienia silnych jednostek SI zostały poznane i możliwe do wdrożenia.

W celu poradzenia sobie z tą potencjalną sytuacją kluczowym celem przejściowych technik redukcji ryzyka SI jest umożliwienie zastosowania superinteligencji przy jednoczesnym zminimalizowaniu ryzyka jednostek SI. W zakresie, w jakim zarządzanie ryzykiem przejściowej SI może opóźnić to ryzyko, zapewniając bezpieczny dostęp do potężnych inteligentnych zasobów, może to się przyczynić do rozwiązania bardziej fundamentalnych problemów na kilka sposobów: przez zwiększenie czasu na dalsze badania, przez informowanie badaczy o konkretnych doświadczeniach oraz, być może, umożliwiając nam użycie superinteligencji rozwiązującej problemy do rozwiązania kwestii superinteligentnych jednostek.

W celu zajęcia się ryzykiem w referencyjnej problematycznej sytuacji można wykorzystać zdolności usprawniania superinteligentnej SI do zadania produkowania (destylacji) najprostszych możliwych uczniów ogólnego przeznaczenia, definiując prostotę za pomocą odpowiedniego wskaźnika minimalnej długości opisu. Można z dużym prawdopodobieństwem założyć, że uczniowie po destylacji MDL opracowani z wykorzystaniem odpowiednich protokołów nie będą mieć znacznej specyficznej wiedzy w obszarach niezwiązanych bezpośrednio z pomyślnym wykonywaniem zestawu kryteriów uczenia.

Niepewności dotyczące zawartości wciąż nieprzezroczystych systemów sztucznej inteligencji (według konserwatywnej hipotezy) można ograniczyć, ucząc wiele instancji uczniów destylowanych MDL z ukierunkowaną, skontrolowaną wiedzą obejmującą programy nauczania dla określonych specjalności. Wtórna destylacja może dodatkowo zawęzić zdobytą wiedzę do tej zasadniczej potrzebnej do wykonania zadań specyficznych dla danej dziedziny, ale jakościowo superinteligentnych. Protokoły punktu kontrolnego/restartu, jako kolejny środek pomiaru wiedzy, mogą wykluczać kumulatywne uczenie się z kolejnych strumieni zadań.

W końcu, główne ograniczenia wąskiej specjalizacji dziedzinowej mogą zostać zniwelowane przez łączenie możliwości wąskich specjalizacji w celu stworzenia bardziej kompleksowych systemów. Odpowiednie architektury mogą umożliwić systemom rozwiązywanie problemów, które obejmują komunikację z użytkownikami ludzkimi przy jednoczesnym ograniczeniu ogólnej wiedzy o świecie. W związku z tym w poniższym aneksie opisano bardziej szczegółowo potencjalną architekturę interaktywnej inżynierii z obsługą SI.

**Tabela 6.6 Zestaw możliwych do zastosowania technik przejściowego zarządzania ryzykiem SI**

| | |
|---|---|
| Destylacja inteligencji | Aby kontrolować początkową ilość informacji |
| Pomiar wiedzy | Aby kontrolować wprowadzanie informacji |
| Punkt kontrolny/restart | Aby kontrolować przechowywanie informacji |
| Ukierunkowane programy nauczania | Aby szkolić specjalistów z wąskich dziedzin |
| Architektury modułowe | Aby komponować specjalistów do praktycznych zadań |

W tabeli 6.6 podsumowano zestaw technik, które złożone kreatywnie i starannie mogą zapewnić znakomite podejście do kształtowania treści i funkcjonalnych możliwości superinteligentnych systemów SI.

Powyższe techniki same w sobie nie mogą zapewnić bezpieczeństwa, ponieważ modułowy skład specjalistycznych systemów SI mógłby zostać wykorzystany do wdrożenia systemów o dosadnych i tak naprawdę nieograniczonych superinteligentnych możliwościach. Pomimo że kryteria dla niezawodnie bezpiecznych zastosowań SI nie są jeszcze dobrze poznane, można jednak oczekiwać, że dobrze wybrane strategie wykorzystujące te techniki mogą znacznie rozszerzyć zakres rozpoznawalnie bezpiecznego obszaru zastosowań.

Wreszcie, wychodząc poza przyrostowe rozszerzania bezpiecznych zastosowań SI, być może najważniejszą motywacją do kontynuowania tej linii badań jest możliwość, że strategie bezpiecznego stosowania superinteligentnych mechanizmów rozwiązywania problemów mogłyby być wskazówką dotyczącą strategii stosowania superinteligencji do rozwiązywania fundamentalnych problemów dotyczących superinteligentnych jednostek.

## PODZIĘKOWANIA

W pracy wykorzystano rezultaty rozmów prowadzonych z wieloma członkami i podmiotami stowarzyszonymi z Future of Humanity Institute, w tym z: Nickiem Bostromem, Stuartem Armstrongiem, Owenem Cotton-Barrattem, Paulem Christiano, Danielem Deweyem, Toby Ordem i Andersem Sandbergiem, a także w latach powstawania pracy z Markiem Millerem pracującym obecnie w Google Research oraz z moim ówczesnym doradcą MIT, Marvinem Minskim.

## ANEKS: BEZPIECZNE ARCHITEKTURY DLA SUPERINTELIGENTNEJ INŻYNIERII

Superinteligentna inżynieria oparta na SI jest ważna nie tylko ze względu na jej potencjalne zastosowania, ale także jako przykład, w którym role specjalizacji, modułowości i kompozycji zadań są silne i stosunkowo dobrze zrozumiane.

Wysoce funkcjonalne systemy inżynieryjne z obsługą SI powinny:

- Omawiać wymagania projektowe z użytkownikami.
- Generować projekty kandydatów.
- Testować projekty kandydatów podczas symulacji.
- Oceniać wydajność projektu.
- Prezentować i wyjaśniać projekty użytkownikom.
- W razie potrzeby powtarzać cykle projektowe.
- Pamiętać o odkryciach projektowych.

Przedstawiona tutaj architektura sugeruje, w jaki sposób powyższe możliwości można bezpiecznie zapewnić za pomocą modułowej kompozycji specjalistów, oraz odpowiednio opisuje dekompozycję zadań, która umożliwiłby interakcję użytkownika, iteracyjne projektowanie i ocenę oraz kumulatywne, specyficzne dla dziedziny, uczenie się (w efekcie zapamiętywanie).

Na rysunku 6.5 przedstawiono uproszczony schemat proponowanego rozkładu zadań i powiązanych interfejsów.

### *A.1. Podsystem człowiek – interfejs*

W tej koncepcji podsystem człowiek – interfejs składa się z dwóch warstw specjalistów: zewnętrzna część to „*chat-sketchbot*", który służy jako inteligentny, interaktywny interfejs człowieka o połączonych kompetencjach dotyczących języka i diagramów związanych z dziedziną. W podsystemie zawarta jest wiedza wyspecjalizowana w odniesieniu do dziedzin inżynierii, języka użytkownika, ustawień preferencji itp. Interfejs skierowany do wewnątrz *chat-sketch* bota tworzy adnotowane diagramy, tabele kryteriów wydajności i tym podobne.

Następnie diagramy, tabele itp. są przekazywane do „generatora specyfikacji", który tworzy formalny i zasadniczo fizyczny opis zadania inżynierskiego. Odwrotny „generator objaśnień" tłumaczy fizyczne opisy na formy, które zewnętrzny specjalista może przedstawić użytkownikowi. Podczas iteracyjnej specyfikacji zadania generator objaśnień może zgłaszać wymagania, które generator specyfikacji oznaczył jako niejednoznaczne, niespójne lub niemożliwe do spełnienia.

Taki wieloskładnikowy podsystem człowiek – interfejs sam z siebie nie odgrywa żadnej roli w zadaniach inżynierskich. Kompetencje inżynierskie systemu zależą od zawartości wewnętrznej skrzynki na rysunku 6.5.

### *A.2. Podsystem wyspecjalizowanej inżynierii*

Mechanizmy rozwiązywania problemów inżynieryjnych mogą zostać rozłożone na generatory propozycji projektów oraz ewaluatory tych propozycji. Te ostatnie komponenty mogą testować i oceniać projekty pod względem ograniczeń fizycznych, kryteriów i wskaźników wydajności.

Na rysunku 6.5 przedstawiono schemat systemu na tym poziomie abstrakcji, w tym potencjał generowania procesów z katalogów wcześniejszych rozwiązań problemów projektowych oraz możliwość wykorzystania zewnętrznych specjalistów w zakresie modelowania fizycznego i symulacji, a także przekazywanie od czasu do czasu projektów do opiekuna katalogu, który przechowuje i indeksuje projekty spełniające kryteria nowości i wydajności.

Należy pamiętać, że przechowywanie projektów w postaci kanonicznych, sparametryzowanych reprezentacji MDL może nie tylko zmniejszyć ich zawartość informacyjną, ale zwykle rozszerza ich ogólność zastosowania i ułatwia wyszukiwanie.

Umożliwiając przechowywanie i wyszukiwanie projektów za pośrednictwem katalogu, można wdrożyć skuteczną, a jednocześnie ściśle specyficzną dla dziedziny formę kumulatywnego uczenia się. W efekcie przechowywanie i wyszukiwanie za pośrednictwem katalogu umożliwia systemom naukę i udostępnianie rosnącego zestawu tymczasowych reguł „jeśli – to" w zakresie projektowania inżynierskiego. Alternatywnie przechowywanie i wyszukiwanie można postrzegać jako formę zapamiętywania.

### *A.3. Architektura systemów*

Na rysunku 6.5 przedstawiono schematy systemów inżynieryjnych o wysokim poziomie abstrakcji i agregacji. W praktyce system inżynieryjny byłby wdrażany jako bardziej szczegółowa sieć podsystemów. W inżynierii na ogół forma podąża za funkcją, zarówno w projektowanych produktach, jak i w procesach projektowych. Byłoby zatem naturalne tworzenie szablonów procesów i architektur systemów inżynieryjnych na znanych wzorcach specjalizacji zadań w organizacjach inżynierskich.

Organizacja zadań w inżynierii, dla wszystkich oprócz prostych lub powtarzalnych zadań, obejmuje odgórny, hierarchiczny rozkład wymagań całego systemu na wymagania podsystemów oraz w niższych poziomach rozkład zadań projektowych na konkretne specjalizacje, takie jak optyka, konstrukcje, elektronika itp. Funkcjonalnie każda relacja w takiej organizacji pociąga za sobą iteracyjną, dwukierunkową wymianę reprezentatywnych dla dziedziny zadań i proponowanych rozwiązań, ponieważ iteracyjne generowanie i ocena projektu są charakterystyczne dla inżynieryjnych zadań projektowych.

Znacznie więcej można powiedzieć o potencjalnych architekturach i zastosowaniach systemów inżynieryjnych opartych na sztucznej inteligencji, ale powyższy opis oddaje poczucie abstrakcyjnych związków między strukturami zadań, specjalizacją i uczeniem się.

Charakter specjalistycznych ról w inżynierii może dać bardziej konkretne pojęcie o tym, w jaki sposób uczniowie MDL mogą być wykorzystywani do tworzenia specjalistów przez kształcenie instancji uczących z ukierunkowaną wiedzą i zadaniami w danej dziedzinie, a następnie do destylacji wtórnej specyficznej dla danej dziedziny.

### *A.4. Względy bezpieczeństwa i uogólnienia*

Biorąc pod uwagę architekturę opisaną powyżej, wydaje się, że istnieje dobry powód, aby sądzić, że techniki destylacji inteligencji, specjalizacji i modułowości architektonicznej mogłyby umożliwić opracowanie i stosowanie szeregu systemów inżynieryjnych działających na poziomie superinteligentnym w bezpieczny sposób, bez ponoszenia znacznego ryzyka związanego z silnymi jednostkami SI.

Koneserzy subtelnych mechanizmów ryzyka SI rozpoznają, że systemy opracowane i zastosowane w *formalny* sposób zgodnie z przedstawionym powyżej szablonem mogą jednak stanowić niedopuszczalne ryzyko wewnętrzne. Tryby środki – cele tworzą kontinuum, a postrzegane abstrakcyjnie to kontinuum obejmuje zarówno projektowanie obwodów, jak i planowanie strategiczne.

Z tego samego powodu może być jednak owocne badanie uogólnień superinteligentnych systemów inżynieryjnych, prowadzące dokładniejszą analizę potencjalnych architektur, aplikacji, ryzyka i środków przeciwdziałania ryzyku. Można oczekiwać, że strategie wykorzystujące wyspecjalizowane architektury rozwiązywania problemów uogólnią się na szeroki zakres zadań SI, a zrozumienie zakresu tych strategii może potencjalnie przyczynić się do rozwiązania odpowiednio szerokiego zakresu problemów obejmujących bezpieczne zastosowanie superinteligentnych funkcji rozwiązywania problemów.

K. Eric Drexler