Poziomy organizacji w inteligencji ogólnej

Sekcja 1 omawia podstawy koncepcyjne inteligencji ogólnej jako dyscypliny, orientując ją w Zintegrowanym Modelu Przyczynowym Tooby'ego i Cosmidesa; Sekcja 2 stanowi główną część artykułu i omawia funkcjonalny rozkład inteligencji ogólnej na złożony supersystem współzależnych wewnętrznie wyspecjalizowanych procesów i strukturyzuje opis, wykorzystując pięć kolejnych poziomów organizacji funkcjonalnej: kod, modalności sensoryczne, koncepcje, myśli i rozważania. Sekcja 3 omawia prawdopodobne różnice między ludźmi a SI i wskazuje na kilka podstawowych zalet, które umysły-w-ogóle-potencjalnie posiadają w porównaniu z obecnymi rozwiniętymi inteligencjami, zwłaszcza w odniesieniu do rekurencyjnego samodoskonalenia.

Podstawy inteligencji ogólnej

Czym jest inteligencja? U ludzi – obecnie jedynych znanych inteligentnych bytów – inteligencja to mózg ze stu miliardami neuronów i stu bilionami synaps; mózg, w którym sama kora mózgowa jest zorganizowana w 52 cytoarchitektonicznie odrębne obszary na półkulę. Inteligencja nie jest złożonym wyrazem prostej zasady; inteligencja jest złożonym wyrazem złożonego zestawu zasad. Inteligencja to supersystem złożony z wielu wzajemnie zależnych podsystemów – podsystemów wyspecjalizowanych nie tylko w określonych umiejętnościach środowiskowych, ale także w określonych funkcjach wewnętrznych. Serce nie jest wyspecjalizowanym organem, który umożliwia nam ściganie ofiary; serce jest wyspecjalizowanym organem, który dostarcza tlen do organizmu. Usuń serce, a rezultatem nie będzie mniej wydajny człowiek ani mniej wyspecjalizowany człowiek; rezultatem jest system, który przestaje działać. Dlaczego inteligencja? Przyczyną ludzkiej inteligencji jest ewolucja – działanie doboru naturalnego na populacji genetycznej, w której organizmy rozmnażają się w sposób różnicowy w zależności od dziedzicznej zmienności cech. Inteligencja jest ewolucyjną zaletą, ponieważ pozwala nam modelować, przewidywać i manipulować rzeczywistością. Problemy ewolucyjne nie ograniczają się do stereotypowych kontekstów przodków, takich jak uciekające lwy czy trzaskające włócznie; nasza inteligencja obejmuje zdolność do modelowania rzeczywistości społecznych składających się z innych ludzi oraz zdolność do przewidywania i manipulowania wewnętrzną rzeczywistością umysłu. Filozofowie umysłu czasami definiują „wiedzę" jako wzorce poznawcze, które mapują rzeczywistość zewnętrzną [76], ale mapowanie powierzchniowe nie ma wrodzonej użyteczności ewolucyjnej. Inteligencja wymaga czegoś więcej niż biernej korespondencji między wewnętrznymi reprezentacjami a danymi sensorycznymi lub między danymi sensorycznymi a rzeczywistością. Poznanie wykracza poza bierne denotacje; może przewidywać przyszłe dane sensoryczne na podstawie przeszłych doświadczeń. Inteligencja wymaga korespondencji wystarczająco silnej, aby organizm mógł wybierać między przyszłościami, wybierając działania na podstawie ich przyszłych rezultatów. Inteligencja w pełni ludzkim sensie wymaga zdolności do manipulowania światem poprzez rozumowanie wstecz od mentalnego obrazu pożądanego wyniku w celu stworzenia mentalnego obrazu koniecznych działań. (Wcześniej te wstępne testy zdolności są sformalizowane jako sensoryczne, predykcyjne, decydujące i manipulacyjne powiązania między modelem a punktem odniesienia.) Zrozumienie ewolucji ludzkiego umysłu wymaga czegoś więcej niż klasycznego darwinizmu; wymaga współczesnego „neo-darwinowskiego" lub „genetyki populacyjnej" rozumienia ewolucji – Zintegrowanego Modelu Przyczynowego przedstawionego przez [98]. Jednym z najważniejszych pojęć w ICM jest „złożona adaptacja funkcjonalna". Adaptacje ewolucyjne są napędzane przez naciski selekcyjne działające na geny. Wkład danego genu w sprawność jest określany przez regularności całego środowiska, w tym zarówno środowiska zewnętrznego, jak i środowiska genetycznego. Adaptacja zachodzi w odpowiedzi na statystycznie obecną złożoność genetyczną, a nie tylko statystycznie obecne konteksty środowiskowe. Nowa adaptacja wymagająca obecności poprzedniej adaptacji nie może się rozprzestrzeniać, chyba że wymagana adaptacja jest obecna w środowisku genetycznym z

wystarczającą regularnością statystyczną, aby uczynić nową adaptację powtarzającą się ewolucyjną przewagą. Ewolucja wykorzystuje istniejącą złożoność genetyczną do budowania nowej złożoności genetycznej, ale ewolucja nie wykazuje żadnej dalekowzroczności. Ewolucja nie tworzy złożoności genetycznej, chyba że jest to natychmiastowa korzyść, a to jest fundamentalne ograniczenie w opisach ewolucji złożonych systemów. Złożone adaptacje funkcjonalne – adaptacje, które wymagają wielu cech genetycznych, aby zbudować złożony, współzależny system w fenotypie – są zazwyczaj i koniecznie uniwersalne w obrębie gatunku. Niezależna wariancja w każdym z genów tworzących złożony, współzależny system szybko zredukowałaby do nieistotności prawdopodobieństwo, że jakikolwiek fenotyp będzie posiadał w pełni funkcjonujący system. Aby podać przykład w uproszczonym świecie, gdyby niezależne geny dla „siatkówki", „soczewki", „rogówki", „tęczówki" i „nerwu wzrokowego" miałyby niezależną częstość 20% w populacji genetycznej, prawdopodobieństwo przypadkowego urodzenia się jakiejkolwiek jednostki z kompletną gałką oczną wynosiłoby 3125:1. Dobór naturalny, żywiąc się zmiennością, zużywa ją . Większość złożoności genetycznej w dowolnym pojedynczym organizmie składa się z głębokiego puli złożonych adaptacji funkcjonalnych pangatunkowych, przy czym naciski selekcyjne działają na powierzchniową pianę indywidualnych zmienności. Celem sztucznej inteligencji nie jest powierzchowna zmienność, która sprawia, że jeden człowiek jest nieco mądrzejszy od drugiego, ale raczej ogromny zasób złożoności, który oddziela człowieka od ameby. Musimy unikać rozpraszania się przez powierzchowne zróżnicowania, które zajmują cały nasz codzienny wszechświat społeczny. Różnice między ludźmi to punkty, w których rywalizujemy i cechy, których używamy do rozpoznawania naszych towarzyszy, a zatem łatwo jest popaść w poświęcanie im zbyt dużej uwagi. Jeszcze większym problemem dla potencjalnych analityków panludzkiej złożoności jest to, że podstawy umysłu nie są otwarte na introspekcję. Postrzegamy tylko najwyższe poziomy organizacji umysłu. Możesz pamiętać przyjęcie urodzinowe, ale nie możesz pamiętać swojego hipokampa kodującego pamięć. Czy introspekcja lub argument ewolucyjny są istotne dla AI? W jakim stopniu prawdy o ludziach mogą być wykorzystane do przewidywania prawd o AI i w jakim stopniu wiedza o ludziach pozwala nam tworzyć projekty AI? Jeśli jedynym celem AI jako dziedziny badań jest testowanie teorii na temat ludzkiego poznania, to tylko prawdy o ludzkim poznaniu są istotne. Ale chociaż ludzka kognitywistyka stanowi uzasadniony cel, nie jest to jedyny powód, aby zajmować się AI; można również zajmować się AI jako celem samym w sobie, w przekonaniu, że AI będzie użyteczna i korzystna. Z tej perspektywy liczy się jakość uzyskanej inteligencji, a nie środki, za pomocą których jest ona osiągana. Jednak właściwe wykorzystanie tego egalitarnego punktu widzenia należy odróżnić od historycznego wykorzystania „techniki przynęty i podmiany", w której „inteligentna SI" jest redefiniowana od jej intuicyjnego znaczenia „SI jako rozpoznawalnej osoby", jednocześnie z prezentacją projektu SI, który pomija większość funkcjonalnych elementów ludzkiej inteligencji i nie oferuje ich zamienników. Istnieje różnica między złagodzeniem ograniczeń dotyczących środków, za pomocą których „inteligencja" może być dopuszczalnie osiągnięta, a obniżeniem standardów, według których oceniamy wyniki jako „inteligencję". Można zatem odejść od metod przyjętych przez ewolucję, ale czy jest to mądre? Ewolucja często znajduje dobre sposoby, ale rzadko najlepsze sposoby. Ewolucja jest użyteczną inspiracją, ale niebezpiecznym szablonem. Ewolucja jest dobrym nauczycielem, ale od nas zależy, czy mądrze zastosujemy lekcje. Ludzie nie są dobrymi przykładami umysłów w ogólności; ludzie są gatunkiem ewoluującym z architekturą poznawczą i emocjonalną dostosowaną do kontekstów łowców-zbieraczy i procesów poznawczych dostrojonych do działania na podłożu masowo równoległych 200-hercowych neuronów biologicznych. Ludzie zostali stworzeni przez ewolucję, proces nieinteligentny; sztuczna inteligencja zostanie stworzona przez inteligentne procesy, którymi są ludzie. Ponieważ ewolucji brakuje przewidywania, złożone funkcje nie mogą ewoluować, chyba że ich przesłankami są ewolucyjne korzyści z innych powodów. Ludzka linia ewolucyjna nie ewoluowała w kierunku ogólnej inteligencji; raczej linia hominidów ewoluowała inteligentniejsze i bardziej złożone systemy, którym brakowało ogólnej inteligencji, aż w końcu skumulowany magazyn istniejącej

złożoności zawierał wszystkie narzędzia i podsystemy potrzebne ewolucji, aby natknąć się na ogólną inteligencję. Nawet to jest zbyt antropocentryczne; powinniśmy raczej powiedzieć, że ewolucja naczelnych natknęła się na gradient sprawności, którego ścieżka obejmuje podgatunek Homo sapiens sapiens, który podgatunek wykazuje jeden szczególny rodzaj ogólnej inteligencji. Eliezer Yudkowsky Ludzcy projektanci sztucznej inteligencji, w przeciwieństwie do ewolucji, będą posiadać zdolność planowania z wyprzedzeniem ogólnej inteligencji. Co więcej, w przeciwieństwie do ewolucji, ludzki planista może przeskakiwać ostre gradienty sprawności, wykonując wiele równoczesnych działań; ludzki projektant może wykorzystać przewidywanie, aby zaplanować wiele nowych komponentów systemu jako część skoordynowanej aktualizacji. Człowiek może podejmować obecne działania w oparciu o przewidywaną zgodność z przyszłymi planami. Tak więc ontogeneza SI nie musi powtarzać ludzkiej filogenezy. Ponieważ ewolucja nie może natknąć się na wielkie projekty supersystemów, dopóki podsystemy nie rozwiną się z innych powodów, filogeneza linii ludzkiej charakteryzuje się rozwojem od bardzo złożonej inteligencji nieogólnej do bardzo złożonej inteligencji ogólnej poprzez warstwową akrecję adaptacyjnej złożoności leżącej w kolejnych poziomach organizacji. Z kolei celowo zaprojektowana SI prawdopodobnie zacznie jako zestaw podsystemów w stosunkowo prymitywnym i nierozwiniętym stanie, ale mimo to już zaprojektowanym w celu utworzenia funkcjonującego supersystemu1. Ponieważ ludzka inteligencja jest ewolucyjnie niedawna, ogromna większość złożoności tworzącej człowieka ewoluowała w nieobecności ogólnej inteligencji; reszta systemu nie miała jeszcze czasu na adaptację. Gdy supersystem AI posiada jakikolwiek stopień inteligencji, bez względu na to, jak prymitywny, inteligencja ta staje się narzędziem, którego można użyć do budowy dalszej złożoności. Podczas gdy linia ludzka rozwinęła się z bardzo złożonej inteligencji nieogólnej w bardzo złożoną inteligencję ogólną, udany projekt AI ma większe szanse rozwinąć się z prymitywnej inteligencji ogólnej w złożoną inteligencję ogólną. Należy zauważyć, że prymitywny nie oznacza architektonicznie prosty. Właściwy zestaw podsystemów, nawet w stanie prymitywnym i uproszczonym, może funkcjonować razem jako kompletny, ale głupi umysł, który następnie zapewnia ramy dla dalszego rozwoju. Nie oznacza to, że SI można sprowadzić do pojedynczego algorytmu zawierającego „esencję inteligencji". Supersystem poznawczy może być „prymitywny" w stosunku do człowieka i nadal wymagać ogromnej ilości złożoności funkcjonalnej. Przyznaję, że jestem stronniczy wobec poszukiwań pojedynczej esencji inteligencji; wierzę, że poszukiwanie pojedynczej esencji inteligencji leży u podstaw poprzednich niepowodzeń SI. Prostota jest Graalem fizyki, nie SI. Fizycy zdobywają Nagrody Nobla, gdy odkrywają wcześniej nieznaną warstwę bazową i wyjaśniają jej zachowania. Wiemy już, jak wygląda ostateczna dolna warstwa sztucznej inteligencji; wygląda jak jedynki i zera. Naszym zadaniem jest zbudowanie czegoś interesującego z tych jedynek i zer. Formalizm Turinga nie rozwiązuje tego problemu bardziej niż elektrodynamika kwantowa mówi nam, jak zbudować rower; wiedza o abstrakcyjnym fakcie, że rower jest zbudowany z atomów, nie mówi, jak zbudować rower z atomów – jakich atomów użyć i gdzie je umieścić. Podobnie abstrakcyjna wiedza, że neurony biologiczne implementują ludzką inteligencję, nie wyjaśnia ludzkiej inteligencji. Klasyczny szum wokół wczesnych sieci neuronowych, że używały „tej samej równoległej architektury, co mózg człowieka", powinien być co najwyżej twierdzeniem o używaniu tej samej równoległej architektury, co mózg dżdżownicy. Nauka o zrozumieniu organizacji życia jest bardzo różna od fizyki lub chemii, gdzie parsymonizm ma sens jako kryterium teoretyczne. Badanie organizmów jest bardziej jak inżynieria wsteczna, w której można mieć do czynienia z dużą liczbą bardzo różnych komponentów, których heterogeniczna organizacja jest wyjaśniana przez sposób, w jaki oddziałują one na siebie, aby wytworzyć funkcjonalny wynik. Ewolucja, konstruktor organizmów żywych, nie ma uprzywilejowanej tendencji do budowania w projektach zasad działania, które są proste i ogólne. Dziedzina sztucznej inteligencji cierpi z powodu dużej, długotrwałej dawki generyczności i koncepcji czarnej skrzynki, czystej karty, tabula rasa, przenikających ze Standardowego Modelu Nauk Społecznych (SSSM) . Ogólny projekt wyzwolenia AI ze szponów SSSM to więcej pracy, niż chciałbym podjąć w tym artykule, ale

jednym z problemów, z którym należy się natychmiast uporać, jest zazdrość o fizykę. Rozwój fizyki w ciągu ostatnich kilku stuleci charakteryzował się odkryciem ujednoliconych równań, które zgrabnie leżą u podstaw wielu złożonych zjawisk. Większość ostatnich pięćdziesięciu lat w AI można opisać jako poszukiwanie podobnej ujednoliconej zasady, która, jak się uważa, leży u podstaw złożonego zjawiska inteligencji. Zazdrość o fizykę w AI to poszukiwanie pojedynczego, prostego procesu leżącego u podstaw, z oczekiwaniem, że to jedno odkrycie odsłoni wszystkie sekrety inteligencji. Tendencja do traktowania nowych podejść do AI tak, jakby były nowymi teoriami fizyki, może przynajmniej częściowo wyjaśniać historię AI pełną nadmiernych obietnic i nadmiernego uproszczenia. Przypisywanie całej ogromnej funkcjonalności ludzkiej inteligencji jednemu opisowemu aspektowi – że mózgi są „równoległe", „rozproszone" lub „stochastyczne"; że umysły stosują „dedukcję" lub „indukcję" – skutkuje porażką (przesadnie rozdmuchaną porażką), ponieważ projekt obiecuje, że cała funkcjonalność ludzkiej inteligencji wyślizgnie się z jakiejś prostej zasady. Efekty zazdrości o fizykę mogą być bardziej subtelne; pojawiają się również w braku interakcji między projektami AI. Zazdrość o fizykę dała początek serii projektów AI, które mogły wykorzystywać tylko jeden pomysł, ponieważ każda nowa hipoteza dotycząca prawdziwej istoty inteligencji była testowana i odrzucana. Programy AM i Eurisko Douglasa Lenata – chociaż wyniki były kontrowersyjne i mogły być lekko przesadzone [85] – mimo to wykorzystywały bardzo intrygujące i fundamentalne wzorce e-design, aby dostarczać znaczące i niespotykane dotąd wyniki. Pomimo tego wzorce projektowe Eurisko, takie jak samomodyfikujące się rozkładalne heurystyki, nie były prawie w ogóle ponownie wykorzystywane w późniejszych AI. Nawet późniejszy projekt Cyc Lenata [62] najwyraźniej nie wykorzystuje ponownie pomysłów opracowanych w Eurisko. Z perspektywy współczesnego programisty, przyzwyczajonego do gromadzenia wzorców projektowych i bibliotek kodu, brak wzajemnego zapłodnienia jest zaskakującą anomalią. Można by pomyśleć, że samooptymalizujące się heurystyki byłyby przydatne jako zewnętrzne narzędzie, np. do dostrajania parametrów, nawet jeśli ogólna architektura poznawcza nie pozwalałaby na wewnętrzne wykorzystanie takich heurystyk. Zachowanie pola AI i samego Lenata jest bardziej zrozumiałe, jeśli założymy, że Eurisko traktowano jako nieudaną hipotezę, a nawet jako konkurencyjną hipotezę, a nie jako przyrostowy sukces lub narzędzie wielokrotnego użytku. Lenat próbował samooptymalizujących się heurystyk, ale nie udało im się uzyskać inteligencji; dalej, do Cyc, kolejnej hipotezy! Najczęstsze paradygmaty tradycyjnej AI – drzewa wyszukiwania, sieci neuronowe, algorytmy genetyczne, obliczenia ewolucyjne, sieci semantyczne – mają wspólną cechę, że można je wdrożyć bez konieczności posiadania zapasu istniejącej wcześniej złożoności. Procesy, które stały się tradycyjne, które zostały ponownie wykorzystane, są narzędziami, które są samodzielne i są natychmiast przydatne. Sieć semantyczna to reprezentacja „wiedzy" tak prosta, że można ją dosłownie zapisać na papierze; zatem projekt AI dodający sieć semantyczną nie musi projektować odpowiednika hipokampa do tworzenia wspomnień ani budować modalności sensorycznej do reprezentowania obrazów mentalnych. Sieci neuronowe i obliczenia ewolucyjne nie są generalnie inteligentne, ale są inteligentne generycznie; można je trenować na dowolnym problemie, który ma wystarczająco płytki gradient sprawności w stosunku do dostępnej mocy obliczeniowej. (Chociaż samomodyfikujące się heurystyki Eurisko prawdopodobnie miały ogólność równą lub przewyższającą te bardziej typowe narzędzia, kod źródłowy nie był otwarty, a projekt systemu był zbyt złożony, aby zbudować go w ciągu popołudnia, więc wzorzec projektowy nie został ponownie wykorzystany – przynajmniej tak przypuszczam). Z wyjątkiem sieci semantycznej, którą uważam za całkowicie bankrutującą, samodzielna natura tradycyjnych procesów może czynić z nich przydatne narzędzia do wspierania początkowych etapów ogólnego supersystemu AI. Jednak samodzielne algorytmy nie zastępują inteligencji i nie są kompletnymi systemami. Ogólność nie jest tym samym, co ogólność. „Zazdrość o fizykę" (próba zastąpienia ludzkiego supersystemu poznawczego pojedynczym procesem lub metodą) należy odróżnić od mniej ambitnej próby oczyszczenia projektu ludzkiego umysłu przy jednoczesnym pozostawieniu nienaruszonej podstawowej architektury. Oczyszczanie jest prawdopodobnie

nieuniknione, gdy zaangażowani są ludzcy programiści, ale mimo to jest to problem, do którego należy podchodzić z najwyższą ostrożnością. Chociaż model ewolucji genetyki populacyjnej dopuszcza wiele teoretycznych powodów, dla których obecność cechy może nie oznaczać adaptacyjności (a tym bardziej optymalności), w praktyce zazwyczaj wygrywają adaptatorzy. Spandrele San Marco mogły nie zostać zbudowane dla dekoracyjnej elegancji , ale nadal podtrzymują dach. Oczyszczanie powinno być przeprowadzane nie z dumą z większej prostoty ludzkiego projektu w porównaniu z projektem ewolucyjnym, ale ze zdrową dawką niepokoju, że pominiemy coś ważnego. Przykład: obecnie uważa się, że ludzie mają modułową adaptację do wizualnego rozpoznawania twarzy, ogólnie utożsamianą z częścią kory skroniowej dolnej, chociaż jest to uproszczenie. Na pierwszy rzut oka to oprogramowanie mózgowe wydaje się być archetypowym przykładem funkcjonalności specyficznej dla człowieka, adaptacją do kontekstu ewolucyjnego bez oczywistego analogu dla wczesnej fazy sztucznej inteligencji. Jednakże [9] zasugerował na podstawie dowodów neuropatologicznych (powiązane deficyty), że oprogramowanie mózgowe do rozpoznawania twarzy jest również odpowiedzialne za uogólnione zadanie zdobywania bardzo precyzyjnej wiedzy w domenie wizualnej; zatem dynamika rozpoznawania twarzy może mieć ogólne znaczenie dla budowniczych modalności sensorycznych. Innym przykładem są same modalności sensoryczne. Jak opisano bardziej szczegółowo w sekcji 2, ludzki supersystem poznawczy jest zbudowany tak, aby wymagał użycia modalności sensorycznych, które pierwotnie wyewoluowaliśmy do innych celów. Jednym z dobrych powodów, dla których ludzki supersystem używa modalności sensorycznych, jest to, że modalności sensoryczne istnieją. Modalności sensoryczne są ewolucyjnie starożytne; istniałyby w pierwotnej lub złożonej formie podczas ewolucji wszystkich wyższych poziomów organizacji. Tkanka nerwowa była już poświęcona modalnościom sensorycznym i nadal zużywałaby ATP, nawet jeśli byłaby nieaktywna, choć w mniejszym tempie. Rozważmy przyrostową naturę adaptacji, tak że na samym początku inteligencji hominidów zaangażowana byłaby tylko bardzo niewielka ilość de novo złożoności; weź pod uwagę, że ewolucja nie ma wrodzonego dążenia do elegancji projektu; weź pod uwagę, że adaptacja jest odpowiedzią na całe środowisko, które obejmuje zarówno środowisko zewnętrzne, jak i środowisko genetyczne – wszystkie te powody są prawdopodobne, aby podejrzewać, że ewolucja przerzuca ciężar obliczeniowy na istniejące wcześniej obwody neuronowe, nawet jeśli ludzki projektant zdecydowałby się na użycie oddzielnego podsystemu. Tak więc nie było z natury absurdalne, aby pierwsi wyznawcy AI próbowali uzyskać ogólną inteligencję, która nie wykorzystywałaby żadnych modalności sensorycznych. Dzisiaj mamy co najmniej jeden powód, aby sądzić, że inteligencja niesensoryczna jest złym podejściem; próbowaliśmy i nie zadziałało. Oczywiście, jest to argument zbyt ogólny – dotyczy on w równym stopniu „próbowaliśmy inteligencji nierozpoznającej twarzy i nie zadziałało" lub nawet „próbowaliśmy inteligencji niedwunożnej i nie zadziałało". Prawdziwa siła argumentu wynika ze szczegółowych hipotez dotyczących funkcjonalnej roli modalności sensorycznej w ogólnej inteligencji. Jednak patrząc wstecz, możemy zidentyfikować co najmniej jeden problem metodologiczny: zamiast identyfikować rolę odgrywaną przez modalności w inteligencji, a następnie próbować „oczyścić" projekt poprzez zastąpienie funkcjonalnej roli odgrywanej przez modalności prostszym procesem, pierwsi badacze AI po prostu założyli, że modalności sensoryczne są nieistotne dla ogólnej inteligencji. Pomijanie kluczowych elementów projektu bez zastąpienia ich innymi na podstawie błędnego przekonania, że nie są one istotne dla ogólnej inteligencji, jest błędem, który wykazuje przerażającą synergię z „zazdrością o fizykę". W skrajnych przypadkach – a większość historycznych przypadków była skrajna – projekt ignoruje wszystko, co dotyczy ludzkiego umysłu, z wyjątkiem jednej cechy (logiki, rozproszonego paralelizmu, niejasności itp.), która jest uważana za „klucz do inteligencji". (W moje bardziej pesymistyczne dni czasami zastanawiam się, czy kolejne mody są jedynym sposobem, w jaki wiedza o danej cesze ludzkiej inteligencji staje się powszechna w AI). Mocno argumentuję za „supersystemami", ale nie wierzę, że „supersystemy" są niezbędnym i wystarczającym kluczem do AI. Ogólna inteligencja wymaga właściwego supersystemu z właściwymi podsystemami poznawczymi,

robiących właściwe rzeczy we właściwy sposób. Ludzie nie są inteligentni z racji bycia „supersystemami", ale z racji bycia szczególnym supersystemem, który implementuje ludzką inteligencję. Podkreślam projekt supersystemu, ponieważ wierzę, że dziedzina AI została sparaliżowana przez niewłaściwy rodzaj prostoty – prostotę, która jako ograniczenie projektowe wyklucza wykonalne projekty dla inteligencji; prostotę, która jako metodologia wyklucza przyrostowy postęp w kierunku zrozumienia ogólnej inteligencji; prostota, która z punktu widzenia czyni większość umysłu niewidzialną, z wyjątkiem jednego aspektu, który jest obecnie promowany jako klucz do SI. Jeśli dążenie do prostoty projektu ma być „uznane za szkodliwe", co powinno je zastąpić? Wierzę, że zamiast prostoty powinniśmy dążyć do wystarczająco złożonych wyjaśnień i użytecznie głębokich projektów. W zwykłym programowaniu nie ma powodu zakładać a priori, że zadanie jest ogromnie duże. W SI regułą powinno być, że problem jest zawsze trudniejszy i głębszy, niż się wydaje, nawet po uwzględnieniu tej reguły. Wiedza o tym, że zadanie jest duże, nie pozwala nam sprostać wyzwaniu, po prostu powiększając lub komplikując nasze projekty; wymagana jest pewna konkretna złożoność, a złożoność lub sama złożoność jest gorsza niż bezużyteczna. Niemniej jednak założenie, że jesteśmy bardziej skłonni do niedoprojektowania niż do przeprojektowania, oznacza inne podejście do projektowania, w którym zwycięstwo nigdy nie jest ogłoszone, a nawet po tym, jak problem wydaje się być rozwiązany, nadal próbujemy go rozwiązać. Gdyby to credo miało zostać podsumowane jednym zdaniem, brzmiałoby ono: „Konieczne, ale niewystarczające". Zgodnie z tym credo należy podkreślić, że myślenie supersystemowe jest tylko częścią większego paradygmatu, a otwarty proces projektowania jest sam w sobie „konieczny, ale niewystarczający". Są to pierwsze kroki w kierunku AI, ale nie jedyne pierwsze kroki i z pewnością nie ostatnie kroki

2 poziomy organizacji w deliberatywnej inteligencji ogólnej

Inteligencja w ludzkim supersystemie poznawczym jest wynikiem wielu procesów poznawczych zachodzących na wielu poziomach organizacji. Jednak stwierdzenie to jest niejasne bez hipotez dotyczących konkretnych poziomów organizacji i konkretnych zjawisk poznawczych. Konkretna teoria przedstawiona w sekcji 2 występuje pod nazwą „deliberatywnej inteligencji ogólnej" (DGI). Ludzki umysł, ze względu na swoje akrecyjne ewolucyjne pochodzenie, ma kilku głównych, odrębnych kandydatów na „środek ciężkości" umysłu. Na przykład układ limbiczny jest ewolucyjnie starożytną częścią mózgu, która obecnie koordynuje działania w wielu innych systemach, które później wyrosły wokół niego. Jednak (ostrożnie) rozważając, jak mógłby wyglądać bardziej przewidujący i mniej akrecyjny projekt inteligencji, stwierdzam, że pojedynczy środek ciężkości wyróżnia się jako mający największą złożoność i wykonujący większość istotnej pracy inteligencji, tak że w sztucznej inteligencji, w jeszcze większym stopniu niż u ludzi, ten środek ciężkości prawdopodobnie stałby się centralnym supersystemem umysłu. Ten środek ciężkości to superproces poznawczy, który ludzie introspekcyjnie obserwują poprzez wewnętrzną narrację – proces, którego działanie odzwierciedlają zdania mentalne, które wewnętrznie „mówimy" i wewnętrznie „słyszymy", myśląc o problemie. Aby uniknąć niezręcznego określenia „strumień świadomości" i nacechowanego słowa „świadomość", ten superproces poznawczy będzie dalej nazywany deliberacją.

Koncepcje: ilustracja zasad

Wybranym przeze mnie punktem wejścia do rozważań są słowa – to znaczy słowa, które mentalnie wypowiadamy i mentalnie słyszymy w naszej wewnętrznej narracji. Weźmy na przykład słowo „żarówka" (lub podobne do słowa wyrażenie „żarówka"). Kiedy widzisz litery tworzące „żarówka", fonemy słowa „żarówka" przepływają przez korę słuchową. Jeśli zadanie umysłowe tego wymaga, wizualny przykład kategorii „żarówka" może zostać przywołany jako obrazy mentalne w korze wzrokowej (i powiązanych obszarach wzrokowych). Niektóre z twoich przeszłych wspomnień i doświadczeń, takie jak przypadkowe rozbicie żarówki i ostrożne zamiatanie ostrych kawałków, mogą

być powiązane z koncepcją „żarówki” lub przechowywane pod nią. „Żarówka” jest powiązana z innymi koncepcjami; w eksperymentach z poznawczym przygotowaniem wykazano, że usłyszenie frazy takiej jak „żarówka” przygotuje skojarzone słowa, takie jak „fluorescencyjny” lub „kruchy”, zwiększając szybkość rozpoznawania lub szybkość reakcji, gdy prezentowane są skojarzone słowa [69]. Koncepcja „żarówki” może działać jako kategoria mentalna; opisuje niektóre odniesienia w postrzeganych doświadczeniach sensorycznych lub wewnętrznych obrazach mentalnych, ale nie inne odniesienia; a spośród opisywanych odniesień niektóre opisuje silnie, a inne tylko słabo. Aby jeszcze bardziej ukazać wewnętrzną złożoność koncepcji „żarówki”, chciałbym przedstawić introspekcyjną ilustrację. Przepraszam wszystkich czytelników akademickich, którzy mają silne uprzedzenia filozoficzne do introspekcji; podkreślam, że ćwiczenie to nie ma na celu udowodnienia teorii, ale raczej wprowadzenia i uziemienia koncepcji, które zostaną omówione bardziej szczegółowo później. To powiedziawszy: zamknij oczy i spróbuj natychmiast (bez świadomego rozumowania) zwizualizować trójkątną żarówkę – teraz. Czy to zrobiłeś? Co zobaczyłeś? Kiedy osobiście wykonałem ten test po raz pierwszy, zobaczyłem żarówkę w kształcie piramidy, z wygładzonymi krawędziami, z żarówką na kwadratowej podstawie. Być może widziałeś żarówkę czworościenną zamiast piramidalnej, albo żarówkę z ostrymi krawędziami zamiast gładkich, albo nawet świetlówkę wygiętą w trójkąt równoboczny. Konkretny wynik jest różny; liczy się proces, którego użyłeś, aby dojść do wyobrażeń mentalnych. Nasze wyobrażenie mentalne „trójkątnej żarówki” intuicyjnie wydaje się być wynikiem nałożenia „trójkątny”, formy przymiotnikowej „trójkąta”, na koncepcję „żarówki”. Innymi słowy, nowy wyobrażenie mentalne trójkątnej żarówki jest najwyraźniej wynikiem połączenia treści sensorycznej dwóch wcześniej istniejących koncepcji. (DGI się zgadza, ale założenie to zasługuje na wyraźne wskazanie). Podobnie, połączenie dwóch koncepcji nie jest kolizją, ale strukturalnym nałożeniem; „trójkątny” jest nałożony na „żarówkę”, a nie „żarówkowy” na „trójkąt”. Strukturalne połączenie dwóch koncepcji jest głównym procesem poznawczym. Podkreślam, że nie mówię o interesującej złożoności, która rzekomo ma być odnaleziona w ogólnym wzorcu relacji między pojęciami; mówię o złożoności, która jest bezpośrednio widoczna w konkretnym przykładzie narzucania „trójkątności” na „żarówkę”. Nie „oddalam”, aby przyjrzeć się ogólnemu terenowi pojęć, ale „przybliżam”, aby przyjrzeć się procesom poznawczym potrzebnym do poradzenia sobie z tym pojedynczym przypadkiem. Konkretny przykład narzucania „trójkątności” na „żarówkę” jest nietrywialnym wyczynem umysłu; „trójkątna żarówka” jest trudniejszą kombinacją pojęć niż „zielona żarówka” lub „trójkątny parking”. Proces mentalny wizualizacji „trójkątnej żarówki” przebiega przez umysł bardzo szybko; możliwe jest dostrzeżenie subiektywnych błysków kombinacji pojęć, ale proces ten nie jest naprawdę otwarty na ludzką introspekcję. Na przykład, gdy po raz pierwszy narzucałem „trójkątny” na „żarówkę”, zgłaszałem krótki subiektywny błysk konfliktu wynikającego z próby narzucenia płaskiego 2-D kształtu „trójkątnego” na 3-D koncepcję „żarówki”. Jednak zanim ten konflikt mógł nastąpić, wydawałoby się konieczne, aby jakiś proces poznawczy wybrał już kształt fasety „trójkątnego” do narzucenia – w przeciwieństwie do, powiedzmy, koloru lub szerokości linii przykładu „trójkąta”, który pojawia się, gdy próbuję zwizualizować „trójkąt” jako taki. Jednak ten początkowy wybór kształtu jako kluczowej fasety nie osiągnął poziomu świadomej uwagi. Mogę zgadnąć, jaki jest ukryty proces selekcji – w tym przypadku, że wcześniejsze doświadczenie z użyciem już „zapamiętało” kształt jako wyraźną stronę koncepcji trójkąta, a koncepcja została wyabstrahowana z bazy doświadczalnej, w której kształt, ale nie kolor, był postrzeganym podobieństwem w grupie doświadczeń. Jednak nie mogę faktycznie dokonać introspekcji tego procesu selekcji. Podobnie, mogłem dostrzec istnienie konfliktu i że był to konflikt wynikający z dwuwymiarowej natury „trójkąta” w porównaniu z trójwymiarową naturą „żarówki”, ale sposób, w jaki konflikt został wykryty, nie jest widoczny w subiektywnym spojrzeniu. A rozwiązanie konfliktu, przekształcenie dwuwymiarowego kształtu trójkąta w trójwymiarowy kształt piramidy, było najwyraźniej natychmiastowe z mojego introspekcyjnego punktu widzenia. Ponownie, mogę zgadnąć, jaki jest ukryty proces – w tym przypadku, że kilku już powiązanych koncepcyjnych sąsiadów „trójkąta”

zostało nałożonych na „żarówkę” równolegle, a najlepiej dopasowanych. Ale nawet jeśli to wyjaśnienie jest poprawne, proces ten nastąpił zbyt szybko, aby można go było dostrzec w bezpośredniej introspekcji. Nie mogę wykluczyć możliwości, że w przejściu od trójkąta do piramidy uczestniczył bardziej złożony, bardziej twórczy proces, chociaż nadal obowiązują podstawowe ograniczenia przetwarzania informacji przez człowieka (ograniczenie prędkości 200 impulsów na sekundę podstawowych neuronów). Nie mogę również wykluczyć możliwości, że istniała unikalna trasa szeregowa od trójkąta do piramidy. Tworzenie rzeczywistego obrazu wzrokowo-przestrzennego piramidalnej żarówki jest prawdopodobnie złożonym procesem wizualnym – takim, który implikuje zdolność modalności wzrokowo-przestrzennej do odwrócenia zwykłego przepływu informacji i wysyłania poleceń z cech wysokiego poziomu do cech niskiego poziomu, zamiast wykrywania cech wysokiego poziomu z cech niskiego poziomu. DGI wysuwa hipotezę, że wizualizacja zachodzi poprzez przepływ od kontrolerów cech wysokiego poziomu do kontrolerów cech niskiego poziomu, tworząc artykułowany obraz mentalny w ramach modalności sensorycznej poprzez wieloetapowy proces, który umożliwia wykrywanie konfliktów na wyższych poziomach przed przejściem do niższych poziomów. Ostateczne wyobrażenia mentalne są widoczne introspektywnie, ale proces, który je tworzy, jest w większości niejasny. Niektórzy teoretycy rzucają wyzwanie introspekcji, aby twierdzić, że nasze wyobrażenia mentalne są czysto abstrakcyjne. Istnieją jednak dowody z neuroanatomii, funkcjonalnego neuroobrazowania, patologii zaburzeń neurologicznych i psychologii poznawczej, które wspierają twierdzenie, że wyobrażenia mentalne są bezpośrednio reprezentowane w modalnościach sensorycznych. Pokazują, że wyobrażenia mentalne mogą tworzyć wizualne obrazy następcze6 podobne do, choć słabsze, obrazów następczych wynikających z rzeczywistego doświadczenia wizualnego. Szacują, że podczas gdy kot ma około 106 włókien od jądra kolankowatego bocznego do kory wzrokowej, w przeciwnym kierunku biegnie około 107 włókien. Obecnie nie istnieje żaden wyjaśniający konsensus dotyczący istnienia masywnych projekcji sprzężenia zwrotnego korowo-wzgórzowego, chociaż istnieje wiele konkurujących teorii; zagadka jest oczywiście interesująca dla badacza AI, który zakłada teorię, że tworzenie nowych obrazów mentalnych jest bardziej intensywne obliczeniowo niż percepcja sensoryczna. Wracając do przykładu „trójkątnej żarówki”: gdy wizualno-przestrzenny obraz piramidalnej żarówki został w pełni wyartykułowany, kolejnym introspektywnym spojrzeniem był konflikt w wizualizacji szklanej piramidy – piramida ma ostre krawędzie, a ostre szkło może przeciąć użytkownika. Oznacza to, że obrazy mentalne miały treść semantyczną (wiedzę o składzie materiałowym piramidalnej żarówki), zaimportowaną z oryginalnej koncepcji „żarówki” i dobrze zintegrowaną z reprezentacją wizualną. Podobnie jak większość współczesnych ludzi, wiem z wczesnych ostrzeżeń rodzicielskich i późniejszego potwierdzenia w prawdziwym życiu, że ostre szkło jest niebezpieczne. Wykryty konflikt został rozwiązany przez nałożenie gładkich krawędzi na szklaną piramidę tworzącą piramidalną żarówkę. Ponownie, najwyraźniej nastąpiło to natychmiast; ponownie, sugeruje się nietrywialną ukrytą złożoność. Aby ująć problem w terminach sugerowanych przez [36], proces wyobraźni musiał posiadać lub stworzyć „pokrętło” regulujące przejście obrazu z ostrych krawędzi do zaokrąglonych, a posiadanie lub stworzenie tego pokrętła jest najciekawszą częścią procesu, a nie wybór jednego pokrętła z wielu. Jeśli „pokrętło” zostało stworzone w locie, oznacza to znacznie wyższy stopień kreatywności systemowej niż wybór spośród wcześniej istniejących opcji. Gdy ostateczny konflikt został rozwiązany przez percepcyjne nałożenie wygładzonych krawędzi, ostateczny obraz mentalny przybrał stabilną formę. Ponownie, w tym przykładzie, wszystkie zdarzenia mentalne wydawały się introspektywnie zdarzać się automatycznie i bez świadomych decyzji z mojej strony; oszacowałbym, że cały proces zajął mniej niż sekundę. W połączeniu pojęć kilka błysków pośrednich etapów przetwarzania może być widocznych jako introspektywne przebłyski – zwłaszcza te konflikty, które pojawiają się na poziomie świadomej uwagi, zanim zostaną automatycznie rozwiązane. Jednak ekstremalna szybkość procesu oznacza, że przebłyski są jeszcze bardziej zawodne niż zwykła introspekcja – gdzie introspekcja jest tradycyjnie uważana za zawodną od samego początku. W

pewnym stopniu jest to sedno ilustracji przedstawionej powyżej; niemal cała wewnętrzna złożoność pojęć jest ukryta przed ludzką introspekcją, a wiele teorii AI (nawet w epoce nowożytnej) próbuje zatem wdrożyć pojęcia na poziomie tokenów, np. „żarówka" jako surowy atom LISP. Ten tradycyjny problem jest powodem, dla którego starannie unikałem używania słowa symbol w powyższym wykładzie. W AI termin „symbol" niesie ze sobą ukryte konotacje dotyczące reprezentacji – że symbol jest nagim atomem LISP, którego domniemane znaczenie wynika z jego relacji do otaczających atomów w sieci semantycznej; lub co najwyżej atom LISP, którego zawartość jest strukturą LISP opartą na ramkach (czyli której zawartość jest inną siecią semantyczną). Nawet próby argumentowania przeciwko założeniom projektowym Good Old-Fashioned AI (GOFAI) są często formułowane w terminach GOFAI; na przykład „problem uziemienia symboli". Wiele dyskusji na temat problemu uziemienia symboli podchodzi do problemu tak, jakby projekt zaczynał się od symboli, a następnie dodawano „uziemienie". W niektórych przypadkach ten punkt widzenia bezpośrednio przełożył się na architektury AI; np. tradycyjna sieć semantyczna jest luźno powiązana z koneksjonistycznym systemem sensomotorycznym. DGI należy do istniejącej tradycji, która pyta nie „Jak uziemiamy nasze sieci semantyczne?", ale raczej „Co jest podstawowym materiałem tworzącym te bogate obiekty wysokiego poziomu, które nazywamy „symbolami"?" .Z tego punktu widzenia, bez odpowiedniego podstawowego „materiału symbolicznego", nie ma symboli; po prostu tokeny LISP-a wyrzeźbione w parodii prawdziwych koncepcji i sprowadzone do nieświętego życia przez sofizmat „nazwanie czyni je tak". Wyobraź sobie modalności sensoryczne jako solidne obiekty z metaforyczną powierzchnią złożoną z warstwowych detektorów cech i ich odwrotnych funkcji jako kontrolerów cech. Metaforyczny „symbolstuff" to wzór, który wchodzi w interakcję z detektorami cech, aby przetestować obecność złożonych wzorców w danych sensorycznych lub odwrotnie, wchodzi w interakcję z kontrolerami cech, aby wytworzyć złożone obrazy mentalne. Symbole łączą się poprzez fasetową kombinację swoich symbolstuff, wykorzystując proces, który można nazwać „holonicznym rozwiązywaniem konfliktów", w którym informacje przepływają od kontrolerów cech wysokiego poziomu do kontrolerów cech niskiego poziomu, a konflikty są wykrywane na każdej warstwie w miarę postępu przepływu. „Holonic" to przydatne słowo do opisania jednoczesnego zastosowania redukcjonizmu i holizmu, w którym pojedyncza jakość jest jednocześnie kombinacją części i częścią większej całości [51]. Na przykład pojedynczy detektor cech może wykorzystywać dane wyjściowe detektorów cech niższego poziomu i działać z kolei jako dane wejściowe detektorów cech wyższego poziomu. Należy zauważyć, że „holoniczny" nie oznacza ścisłej hierarchii, a jedynie ogólny przepływ od wysokiego poziomu do niskiego poziomu. Przepraszam za dodanie kolejnego terminu, „holoniczne rozwiązywanie konfliktów", do przestrzeni nazw już zatłoczonej terminami takimi jak „temperatura obliczeniowa" , „Prägnanz" , „sieci Hopfielda", „propagacja ograniczeń" i wieloma innymi. Holoniczne rozwiązywanie konfliktów z pewnością nie jest całkowicie nowym pomysłem i może być nawet zupełnie nieoryginalne w odniesieniu do każdej cechy, ale kombinacja cech, które chcę opisać, nie do końca odpowiada istniejącemu powszechnemu użyciu żadnego z powyższych terminów. „Holoniczne rozwiązywanie konfliktów" ma na celu przekazanie obrazu procesu, który przepływa seryjnie przez warstwową, holoniczną strukturę percepcji, przy czym wykryte konflikty są rozwiązywane lokalnie lub propagowane na wyższy poziom, z ostatecznym rozwiązaniem, które spełnia. Wiele z powyższych terminów, w ich powszechnym użyciu, odnosi się do iterowanego procesu wyżarzania, który dąży do globalnego minimum. Holoniczne rozwiązywanie konfliktów ma być biologicznie prawdopodobne; tj. aby zapewnić płynny przepływ wizualizacji, który jest obliczeniowo wykonalny dla równoległych, ale ograniczonych prędkością neuronów. Holoniczne rozwiązywanie konfliktów nie jest proponowane jako kompletne rozwiązanie problemów percepcyjnych, ale raczej jako aktywne płótno do interakcji pojęć z obrazami mentalnymi. W terminologii teoretycznej holoniczne rozwiązywanie konfliktów jest strukturalnym frameworkiem, w którym można umieścić określone metody wykrywania i rozwiązywania konfliktów. Holoniczne obrazowanie jest medium artysty, w którym symbole malują

obrazy mentalne, takie jak „trójkątna żarówka”. Konstruktywne ujęcie pojęć i symboli musiałoby dostarczyć:

• Opisu tego, w jaki sposób pojęcie jest spełniane i narzucane referencjom w modalności sensorycznej

• Reprezentacji symboli spełniającej (a), która może zawierać wewnętrzną złożoność potrzebną do fasetowej kombinacji pojęć

• Reprezentacji spełniającej (a) i (b), tak że jest obliczeniowo wykonalna do abstrahowania nowych pojęć przy użyciu doświadczenia sensorycznego jako surowca

To nie jest wyczerpująca lista funkcjonalności pojęć; to tylko trzy najbardziej „interesujące” wyzwania. Te wyzwania są interesujące, ponieważ trudność ich jednoczesnego rozwiązania wydaje się być iloczynem mnożnikowym (a nie addytywnym) trudności ich indywidualnego rozwiązania. Inne wymagania projektowe dla konstruktywnego opisu pojęć obejmowałyby: skojarzenie z pobliskimi pojęciami; superkategorie i podkategorie; przykłady zapisane w pamięci; efekty prototypu i typowości [88]; i wiele innych (patrz np. [57]). Interakcja pojęć z modalnościami i interakcja pojęć ze sobą nawzajem ilustrują to, co uważam za kilka ważnych zasad dotyczących podejścia do AI. Pierwszą zasadą jest zasada wielu poziomów organizacji. Fenotyp człowieka składa się z atomów, cząsteczek, białek, komórek, tkanek, organów, układów organów i wreszcie całego ciała – ośmiu rozróżnialnych warstw organizacji, z których każda kolejna warstwa jest zbudowana nad poprzednią, a każda kolejna warstwa obejmuje ewolucyjną złożoność adaptacyjną. Niektóre użyteczne właściwości wyższego poziomu mogą wyłaniać się naturalnie z zachowań niższego poziomu, ale nie wszystkie; właściwości wyższego poziomu podlegają również presji selekcyjnej na dziedziczną zmienność i opracowanie złożonych adaptacji funkcjonalnych. Postulując wiele poziomów organizacji, nie zakładam, że zachowania wszystkich wyższych warstw wyłaniają się automatycznie z najniższej warstwy. Gdybym miał wybrać jeden błąd, który był najbardziej wyniszczający w AI, byłoby to wdrożenie procesu zbyt bliskiego poziomowi tokena – próba wdrożenia procesu wysokiego poziomu bez wdrożenia podstawowych warstw organizacji. Wiele przysłowiowych patologii AI wynika przynajmniej częściowo z pominięcia niższych poziomów organizacji w projekcie. Weźmy na przykład tę wersję „problemu ramowego” – czasami uważaną również za formę „problemu zdrowego rozsądku” – w której inteligentne rozumowanie wydaje się wymagać znajomości nieskończonej liczby przypadków szczególnych. Rozważmy procesor, który dodaje dwie 32-bitowe liczby. Wyższy poziom składa się z dwóch liczb całkowitych, które są dodawane w celu wytworzenia trzeciej liczby całkowitej. Na niższym poziomie obiekty obliczeniowe nie są postrzegane jako nieprzejrzyste „liczby całkowite”, ale jako uporządkowane struktury 32-bitowe. Gdy procesor wykonuje operację arytmetyczną, dwie struktury 32-bitowe zderzają się ze sobą, zgodnie z pewnymi zasadami, które rządzą lokalnymi interakcjami między bitami, a wynikiem jest nowa struktura 32-bitowa. Teraz rozważmy niedole zespołu badawczego, niemającego wiedzy o podstawowej implementacji procesora, który próbuje stworzyć arytmetyczny „system ekspercki” poprzez kodowanie ogromnej sieci semantycznej zawierającej „wiedzę”, że dwa plus dwa daje cztery, dwadzieścia jeden i szesnaście daje trzydzieści siedem itd. Ta gigantyczna tabela wyszukiwania wymaga osiemnastu miliardów miliardów wpisów do ukończenia. W tym hipotetycznym świecie, w którym nie rozumiemy procesu dodawania na niższym poziomie, możemy sobie wyobrazić „zdroworozsądkowy” problem dodawania; uruchomienie rozproszonych projektów internetowych w celu „zakodowania całej szczegółowej wiedzy niezbędnej do dodawania”; problem ramowy dodawania; filozofie formalnej semantyki, zgodnie z którymi token LISP trzydzieści siedem jest znaczący, ponieważ odnosi się do trzydziestu nawet obiektów w świecie zewnętrznym; zasadę projektowania, zgodnie z którą token trzydzieści siedem nie ma wewnętrznej złożoności i raczej nadaje mu znaczenie jego sieć relacji z innymi tokenami; „problem uziemienia liczby”; pełnych nadziei futurologów twierdzących, że poprzednie projekty mające na celu stworzenie sztucznego dodawania

zakończyły się niepowodzeniem z powodu niewystarczającej mocy obliczeniowej; i tak dalej. Do pewnego stopnia jest to niesprawiedliwa analogia. Nawet jeśli eksperyment myślowy jest zasadniczo poprawny, a opisane nieszczęścia wynikałyby z próby uchwycenia opisu arytmetyki na wysokim poziomie bez implementacji podstawowego niższego poziomu, nie dowodzi to, że analogiczny błąd jest źródłem tych nieszczęść w prawdziwym polu AI. I do pewnego stopnia powyższy opis jest niesprawiedliwy nawet jako eksperyment myślowy; arytmetyczny system ekspercki nie byłby tak bankrutujący jak sieci semantyczne. Regularności w „systemie eksperckim do arytmetyki" byłyby rzeczywiste, zauważalne za pomocą prostych i obliczeniowych środków i mogłyby zostać użyte do wywnioskowania, że arytmetyka była podstawowym procesem, który jest reprezentowany, nawet przez Marsjanina czytającego kod programu bez żadnej wskazówki co do zamierzonego celu systemu. Luka między wyższym poziomem a niższym poziomem nie jest absolutna i nieprzekraczalna, jak w sieciach semantycznych. Arytmetyczny system ekspercki, który pomija jeden poziom organizacji, może być odzyskiwalny. Sieci semantyczne pomijają wiele poziomów organizacji. Pominięcie całego doświadczalnego i sensorycznego uziemienia ludzkich symboli nie pozostawia żadnego surowego materiału do pracy. Gdyby wszystkim tokenom LISP w sieci semantycznej nadano losowe nowe nazwy, nie byłoby sposobu, aby wywnioskować, czy G0025 wcześniej oznaczało hamburgera czy Johnny'ego Carsona. [29] opisuje problem uziemienia symboli wynikający z sieci semantycznych jako podobny do próby nauczenia się chińskiego jako pierwszego języka przy użyciu jedynie słownika chińsko-chińskiego. Wierzę, że wiele (choć nie wszystkie) przypadków „problemu zdrowego rozsądku" lub „problemu ram" wynika z próby przechowywania wszystkich możliwych opisów zachowań wysokiego poziomu, które w umyśle ludzkim są modelowane przez wizualizację niższego poziomu organizacji, z którego te zachowania się wyłaniają. Na przykład [58] podaje przykładową listę „wbudowanych wniosków" wyłaniających się z tego, co identyfikują jako metaforę Źródło-Ścieżka-Cel:

• Jeśli przebyłeś trasę do bieżącej lokalizacji, byłeś we wszystkich poprzednich lokalizacjach na tej trasie.

• Jeśli podróżujesz z A do B i z B do C, to podróżowałeś z A do C.

• Jeśli X i Y podróżują bezpośrednią trasą z A do B, a X mija Y, to X jest dalej od A i bliżej B niż Y. • (etc.)

Ogólna inteligencja z modalnością wizualną nie musi jawnie przechowywać nieskończonej liczby takich stwierdzeń w systemie produkcyjnym dowodzącym twierdzeń. Powyższe stwierdzenia można postrzegać w locie, badając obrazowe obrazy mentalne. Zamiast przechowywać wiedzę o trajektoriach, modalność wizualna w rzeczywistości symuluje zachowanie trajektorii. Modalność wizualna wykorzystuje elementy niskiego poziomu, metaforyczne „piksele" i ich holoniczną strukturę cech, których zachowania lokalnie odpowiadają zachowaniom odniesienia w świecie rzeczywistym. Istnieje mapowanie od reprezentacji do odniesienia, ale jest to mapowanie na niższym poziomie organizacji niż próbują uchwycić tradycyjne sieci semantyczne. Korespondencja ma miejsce na poziomie, na którym 13 jest strukturą 00001101, a nie na poziomie, na którym jest liczbą trzynaście. Czasami spotykam się z pewnym zamieszaniem co do różnicy między modalnością wizualną a mikroteorią widzenia. Niewątpliwie mikroteorie w systemach dowodzących twierdzeń są dobrze znane w AI, chociaż osobiście uważam je za paradygmat o małej wartości, więc pewne zamieszanie jest zrozumiałe. Ale ekstrakcja warstwowa cech w modalności wizualnej – co jest ustalonym faktem neuronauki – jest również bardzo dobrze znana nawet w czystej tradycji informatycznej AI i jest dobrze znana od czasu niezwykle wpływowej książki Davida Marra z 1982 r. Vision [65] i wcześniejszych prac. Aby wyraźnie zaznaczyć różnicę, ludzka kora wzrokowa „wie" o wykrywaniu krawędzi, cieniowaniu, fakturach zakrzywionych powierzchni, dysproporcjach obuocznych, stałości kolorów w świetle naturalnym, ruchu względem płaszczyzny fiksacji itd. Kora wzrokowa nie wie o motylach. W rzeczywistości kora wzrokowa „nie wie" nic; modalność sensoryczna zawiera zachowania odpowiadające niezmiennikom

środowiskowym, a nie wiedzę o prawidłowościach środowiskowych. Ilustruje to drugi najgorszy błąd w sztucznej inteligencji, czyli brak rozróżnienia między rzeczami, które mogą być zaprogramowane na stałe, a rzeczami, których trzeba się nauczyć. Nie jesteśmy zaprogramowani, aby wiedzieć o motylach. Ewolucja wyposażyła nas w obwody wzrokowe, które nadają sens obrazowi sensorycznemu motyla, oraz w systemy rozpoznawania obiektów, które tworzą kategorie wizualne. Kiedy widzimy motyla, jesteśmy w stanie rozpoznać przyszłe motyle jako należące do tego samego rodzaju. Czasami ewolucja omija ten system, aby obdarzyć nas instynktami wzrokowymi, ale stanowi to niewielką część wiedzy wizualnej. Współczesny człowiek rozpoznaje ogromną liczbę kategorii wizualnych, które nie mają odpowiedników w środowisku przodków. Jakie problemy wynikają z braku rozróżnienia między rzeczami, które mogą być zaprogramowane na stałe, a rzeczami, których trzeba się nauczyć? „Zaprogramowanie na stałe tego, czego należy się nauczyć" jest tak powszechnie połączone z „załamaniem poziomów organizacji", że trudno jest uporządkować wynikające z tego patologie. Inżynier systemów eksperckich, oprócz przekonania, że wiedza o motylach może być wstępnie zaprogramowana, prawdopodobnie będzie również wierzył, że wiedza o motylach składa się z tokena motyla LISP, który czerpie swoje znaczenie ze swojej relacji do innych tokenów LISP – zamiast motyla będącego przechowywanym wzorcem, który wchodzi w interakcję z modalnością wizualną i rozpoznaje motyla. Sieć semantyczna nie tylko nie jest bogata, ale także nie ma zdolności do reprezentowania bogactwa. Dlatego przypisałbym problem uziemienia symboli „załamywaniu poziomów organizacji", a nie „wbudowywaniu na stałe tego, czego należy się nauczyć". Ale nawet jeśli programista, który rozumiał poziomy organizacji, próbował ręcznie tworzyć symbole rozpoznające motyle, nadal spodziewałbym się, że powstały wzór motyla nie będzie miał bogactwa wyuczonego wzoru motyla w ludzkim umyśle. Kiedy ludzki układ wzrokowy tworzy wizualną kategorię motyla, nie zapisuje niejasnego, proceduralnego kodeka rozpoznawania motyli, używając abstrakcyjnej wiedzy o motylach, a następnie nie oznacza kodeka na ramce motyla. Ludzka kategoryzacja wizualna abstrahuje kategorię motyla z magazynu wizualnych doświadczeń motyli. Ponadto kategoryzacja wizualna – ogólny proces formowania pojęć, a nie tylko czasowy strumień przetwarzania wizualnego – pozostawia po sobie skojarzenie między pojęciem motyla a przechowywanymi wspomnieniami, z których „motyl" został wyabstrahowany; kojarzy jeden lub więcej przykładów z kategorią motyla; kojarzy kategorię motyla poprzez nakładające się terytorium z innymi kategoriami wizualnymi, takimi jak trzepotanie; tworzy symbole motyla, które mogą łączyć się z innymi symbolami, aby wytworzyć mentalne obrazy niebieskiego motyla; i tak dalej. W stopniu, w jakim człowiekowi brakuje cierpliwości, aby robić te rzeczy, lub w stopniu, w jakim człowiek robi je w kruchy i ręcznie zakodowany sposób, zamiast używać solidnej abstrakcji z chaotycznej bazy doświadczalnej, brak bogactwa będzie skutkiem. Nawet jeśli AI potrzebuje koncepcji stworzonych przez programistę, aby zainicjować dalsze formowanie pojęć, koncepcje bootstrapowe powinny być tworzone przy użyciu wersji narzędziowych sterowanych przez programistę odpowiadających im podsystemów AI, a koncepcje bootstrapowe powinny być zastępowane koncepcjami utworzonymi przez AI tak wcześnie, jak to możliwe. Dwa inne potencjalne problemy wynikające z użycia treści stworzonych przez programistów to nieprzejrzystość i izolacja. Przejrzystość odnosi się do potencjalnej niezdolności podsystemów AI do modyfikowania treści, które powstały poza AI. Jeśli programista tworzy treść poznawczą, powinna ona być przynajmniej tego rodzaju treścią, którą AI mogłaby stworzyć samodzielnie; powinna być treścią w formie, którą podsystemy poznawcze AI mogą manipulować. Najlepszym sposobem na zapewnienie, że AI może modyfikować i wykorzystywać wewnętrzną treść, jest zlecenie AI utworzenia treści. Jeśli podsystemy poznawcze SI są wystarczająco silne, aby samodzielnie tworzyć treść, to miejmy nadzieję, że te same podsystemy będą w stanie dodawać do tej treści, manipulować nią, wyginać ją w odpowiedzi na naciski wywierane przez problem itd. To, co SI tworzy, SI może wykorzystać i ulepszyć. Cokolwiek SI osiągnie samodzielnie, jest częścią umysłu SI; SI „posiada" to i nie pożycza tego po prostu od programistów. Jest to zasada, która wykracza daleko poza abstrahowanie pojęć! Izolacja oznacza, że jeśli pojęcie lub część

wiedzy zostanie podana SI na srebrnej tacy, SI może zostać odizolowana od rzeczy, których SI musiałaby się najpierw nauczyć, aby zdobyć tę wiedzę naturalnie, w trakcie budowania kolejnych warstw zrozumienia, aby poradzić sobie z problemami o coraz większej złożoności. Koncepcja może być również odizolowana od innych rzeczy, których SI nauczyłaby się mniej więcej w tym samym czasie, co może oznaczać niedobór przydatnych skojarzeń i poślizgów. Programiści mogą próbować obejść problem izolacji, ucząc wielu podobnych zagadnień w tym samym czasie, ale nie zastąpi to naturalnej ekologii poznania.

Poziomy organizacji w deliberacji

Model deliberacji przedstawiony w tym rozdziale wymaga pięciu odrębnych warstw organizacji, z których każda jest zbudowana na warstwie bazowej.

• Dolna warstwa to kod źródłowy i struktury danych – złożoność, którą programista manipuluje bezpośrednio. Odpowiednią warstwą dla ludzi są neurony i obwody neuronowe.

• Następna warstwa to modalności sensoryczne. U ludzi archetypowymi przykładami modalności sensorycznych są wzrok, słuch, dotyk, smak, węch itd.; implementowane przez obszary wzrokowe, słuchowe itd. W mózgach biologicznych modalności sensoryczne są najbliższe bycia „zaprogramowanymi na stałe"; zazwyczaj obejmują one jasno zdefiniowane etapy przetwarzania informacji i ekstrakcji cech, czasami z pojedynczymi neuronami odgrywającymi jasno zdefiniowane role. Tak więc modalności sensoryczne są jednymi z najlepszych kandydatów do procesów, które mogą być bezpośrednio kodowane przez programistów bez czynienia systemu krystalicznym i kruchym.

• Następna warstwa to koncepcje. Koncepcje (czasami znane również jako „kategorie" lub „symbole") są abstrahowane z naszych doświadczeń. Abstrakcja reifikuje postrzegane podobieństwo w grupie doświadczeń. Po reifikacji, wspólna jakość może być następnie wykorzystana do określenia, czy nowe wyobrażenia mentalne spełniają jakość, a jakość może zostać narzucona na obraz mentalny, zmieniając go. Po abstrakcji koncepcji „czerwony", możemy wziąć obraz mentalny obiektu nie będącego czerwonym (na przykład trawy) i wyobrazić sobie „czerwoną trawę". Koncepcje to wzorce, które łączą się z obrazami sensorycznymi; koncepcje to złożone, elastyczne, wielokrotnego użytku wzorce, które zostały reifikowane i umieszczone w długoterminowym magazynie.

• Następna warstwa to myśli, zbudowane ze struktur koncepcji. Poprzez narzucanie koncepcji w ukierunkowanych seriach, staje się możliwe budowanie złożonych obrazów mentalnych w przestrzeni roboczej dostarczanej przez jedną lub więcej modalności sensorycznych. Archetypowym przykładem myśli jest ludzkie „zdanie" – układ pojęć, przywoływany przez ich symboliczne znaczniki, z wewnętrzną strukturą i ukierunkowaną informacją, którą można zrekonstruować z liniowej serii słów, używając ograniczeń składni, konstruując złożony obraz mentalny, który można wykorzystać w rozumowaniu. Myśli (i odpowiadające im obrazy mentalne) to jednorazowe struktury, zbudowane z wielokrotnego użytku pojęć, które wdrażają nierekurencyjny umysł w nierekurencyjnym świecie.

• Wreszcie, to sekwencje myśli wdrażają rozważania – wyjaśnianie, przewidywanie, planowanie, projektowanie, odkrywanie i inne działania wykorzystywane do rozwiązywania problemów wiedzy w dążeniu do celów w świecie rzeczywistym.

Chociaż pięciowarstwowy model jest centralny dla teorii inteligencji DGI, zasada Koniecznego, ale Niewystarczającego nadal obowiązuje. Projekt AI nie odniesie sukcesu dzięki „wdrożeniu pięciowarstwowego modelu inteligencji, tak jak ludzki mózg". Musi to być właściwych pięć warstw. Muszą to być właściwe modalności, użyte we właściwych koncepcjach, łączące się, aby stworzyć właściwe myśli poszukujące właściwych celów. Pięciowarstwowy model rozważań nie obejmuje

wszystkiego w teorii umysłu DGI, ale obejmuje znaczny obszar i może być rozszerzony poza superproces rozważań, aby zapewnić luźne poczucie, na jakim poziomie organizacji leży każdy proces poznawczy. Obserwacja, że ludzkie ciało składa się z cząsteczek, białek, komórek, tkanek i organów, nie jest kompletnym projektem ludzkiego ciała, ale mimo to ważne jest, aby wiedzieć, czy coś jest organem, czy białkiem. Na przykład krew nie jest prototypową tkanką, ale składa się z komórek i ogólnie mówi się, że zajmuje poziom organizacji tkankowej ludzkiego ciała. Podobnie hipokamp, w swojej roli podsystemu formowania pamięci, nie jest modalnością sensoryczną, ale można powiedzieć, że zajmuje „poziom modalności": jest to brainware (oddzielny, modułowy fragment obwodów neuronowych); leży powyżej poziomu neuronów/kodów; ma charakterystyczny wzór kafelkowania/okablowania jako wynik złożoności genetycznej; oddziałuje jako równy z podsystemami obejmującymi modalności sensoryczne. Uogólnione definicje pięciu poziomów organizacji mogą być następujące:

Poziom kodu, poziom sprzętu: Nie potrzeba żadnej ogólnej definicji, poza tym, że biologicznym odpowiednikiem jest poziom neuronowy lub poziom oprogramowania.

Poziom modalności: Podsystemy, które u ludzi czerpią swoją adaptacyjną złożoność ze specyfikacji genetycznej – lub raczej ze specyfikacji genetycznej początkowego wzoru kafelkowania i algorytmu samoukładania oraz z ekspozycji na niezmienną złożoność środowiskową. Odpowiednikiem AI jest złożoność, która jest znana z góry programiście i która jest bezpośrednio określana przez wysiłki programisty. Pełne systemy na tym poziomie są modułowymi częściami poznawczego supersystemu – jedną z dużej, ale ograniczonej liczby głównych części tworzących umysł. Jeśli dany system jest modalnością sensoryczną lub systemem, który wyraźnie łączy się z modalnościami sensorycznymi i wykonuje zadania związane z modalnością, system można określić jako poziom modalności. Podobnie podsystem lub podproces głównego systemu poziomu modalności lub podrzędna funkcja takiego podsystemu mogą być również określane jako poziom modalności. W przypadku, gdy termin ten jest niewłaściwy, ponieważ podsystem ma niewielki lub żaden związek z modalnościami sensorycznymi, podsystem ten można nazwać oprogramowaniem mózgowym

Poziom konceptualny: Koncepcje to obiekty poznawcze, które są umieszczane w długoterminowym magazynie i ponownie wykorzystywane jako elementy budulcowe myśli. Uogólnieniem dla tego poziomu organizacji jest wyuczona złożoność: treść poznawcza, która pochodzi ze środowiska i jest umieszczana w długoterminowym magazynie, a tym samym staje się częścią stałego rezerwuaru złożoności, z którym sztuczna inteligencja stawia czoła przyszłym problemom. Termin poziom konceptualny może być opcjonalnie stosowany do dowolnej wyuczonej złożoności, która przypomina kategorie; tj. wyuczona złożoność, która oddziałuje z modalnościami sensorycznymi i działa na modalności sensoryczne. Niezależnie od tego, czy są one podobne do konceptów (kwestia ta zostanie rozważona później), inne przykłady wyuczonej złożoności obejmują deklaratywne przekonania i wspomnienia epizodyczne.

Poziom myśli: Myśl to określona struktura symboli kombinatorycznych, która buduje lub zmienia obrazy mentalne. Uogólnialną właściwością myśli jest ich bezpośredniość. Myśli nie są ewolucyjnie/zaprogramowanym oprogramowaniem mózgowym ani długoterminowym rezerwuarem wyuczonej złożoności; myśli są konstruowane z chwili na chwilę. Myśli stanowią historię życia umysłu nierekurencyjnego w nierekurencyjnym wszechświecie. Uogólniony poziom myśli wykracza poza wypowiedzi mentalne w naszym strumieniu świadomości; obejmuje wszystkie główne zdarzenia poznawcze występujące w świecie aktywnych obrazów mentalnych, zwłaszcza zdarzenia, które obejmują strukturowanie kombinatorycznej złożoności poziomu koncepcji. Rozważanie: Który, podobnie jak poziom kodu, nie wymaga uogólnienia. Rozważanie opisuje działania wykonywane przez wzorce myśli. Wzory w rozwadze nie są tylko epifenomenalnymi właściwościami sekwencji myśli;

poziom rozwagi jest kompletną warstwą organizacji, ze złożonością specyficzną dla tej warstwy. W rozwadze AI to wzorce myśli planują i projektują, przekształcając abstrakcyjne wzorce celów wysokiego poziomu w określone wzorce celów niskiego poziomu; to wzorce myśli rozumują od bieżącej wiedzy do przewidywań dotyczących nieznanych zmiennych lub przyszłych danych sensorycznych; to wzorce myśli rozumują o niewyjaśnionych obserwacjach, aby wymyślać hipotezy dotyczące możliwych przyczyn. Ogólnie rzecz biorąc, rozważanie polega na wykorzystaniu uporządkowanych sekwencji myśli do rozwiązywania problemów wiedzy w celu osiągnięcia celów w świecie rzeczywistym.

Nawet w przypadku uogólnionych poziomów organizacji nie wszystko idealnie pasuje do jednego lub drugiego poziomu. Podczas gdy trichotomia „hardwired-learned-invented" zwykle pasuje do trichotomii „modality-concept-thought", oba są koncepcyjnie odrębne, a czasami korespondencja jest zerwana. Jednak poziomy organizacji są prawie zawsze przydatne – nawet wyjątki od reguły są łatwiej postrzegane jako częściowe odstępstwa niż jako kompletne przypadki szczególne.

Poziom kodu

Poziom kodu składa się z funkcji, klas, modułów, pakietów; typów danych, struktur danych, repozytoriów danych; wszystkich czysto programowych wyzwań tworzenia AI. Sztuczna inteligencja tradycyjnie była znacznie bardziej powiązana z programowaniem komputerowym niż powinna być, głównie z powodu prób nadmiernej kompresji poziomów organizacji i implementacji sekwencji myśli bezpośrednio jako procedur programistycznych lub implementacji koncepcji bezpośrednio jako atomów LISP lub ramek LISP. Poziom kodu leży bezpośrednio pod poziomem modalności lub poziomem oprogramowania mózgowego; przeciek z wyzwań na poziomie modalności może objawiać się jako uzasadnione problemy programistyczne, ale niewiele więcej – nie myśli, treści poznawcze ani metody rozwiązywania problemów na wysokim poziomie. Każdy dobry programista – programista z wyczuciem estetyki – zna nudę rozwiązywania tego samego przypadku szczególnego, raz po raz, w nieco inny sposób; a także triumf myślenia o metaproblemie i tworzenia ogólnego rozwiązania, które rozwiązuje wszystkie przypadki szczególne jednocześnie. Jak zauważa hacker Jargon File, „Prawdziwi hakerzy uogólniają nieciekawe problemy na tyle, by uczynić je interesującymi i je rozwiązać – rozwiązując w ten sposób pierwotny problem jako przypadek szczególny (i, trzeba przyznać, czasami zamieniając kopiec kreta w górę lub górę w płytę tektoniczną)". [82]. Ten idiom nie działa w przypadku ogólnej SI! Prawdziwa SI byłaby ostatecznym ogólnym rozwiązaniem, ponieważ obejmowałaby procesy poznawcze, których ludzcy programiści używają do pisania dowolnego konkretnego fragmentu kodu, ale tego ostatecznego rozwiązania nie można uzyskać poprzez technikę sukcesywnego uogólniania nieciekawych problemów na interesujące. Programowanie to sztuka tłumaczenia ludzkiego modelu mentalnego rozwiązania problemu na program komputerowy; to znaczy sztuka tłumaczenia myśli na kod. Programowanie z natury narusza poziomy organizacji; prowadzi bezpośrednio do pułapek klasycznej SI. Podstawowe procesy niskiego poziomu, które wdrażają inteligencję, mają zasadniczo inny charakter niż sama inteligencja wysokiego poziomu. Kiedy tłumaczymy nasze myśli o problemie na kod, ustanawiamy korespondencję między kodem a treścią wysokiego poziomu naszych umysłów, a nie korespondencję między kodem a dynamicznym procesem ludzkiego umysłu. W zwykłym programowaniu zadaniem jest sprawienie, aby komputer rozwiązał konkretny problem; może to być „interesujący" problem z bardzo dużą domeną, ale nadal będzie to konkretny problem. W zwykłym programowaniu problem rozwiązuje się, biorąc ludzki proces myślowy, który zostałby użyty do rozwiązania instancji problemu, i tłumacząc ten proces myślowy na kod, który może również rozwiązać instancje problemu. Programiści to ludzie, którzy nauczyli się sztuki wymyślania procesów myślowych, zwanych „algorytmami", które polegają wyłącznie na możliwościach, jakie posiada zwykły komputer. Odruchy nabyte przez dobrego, artystycznego programistę stanowią podstawowe niebezpieczeństwo przy rozpoczynaniu ogólnego projektu AI. Programiści są szkoleni w rozwiązywaniu problemów, a

próba stworzenia ogólnej AI oznacza rozwiązanie problemu programistycznego polegającego na stworzeniu umysłu, który rozwiązuje problemy. Istnieje niebezpieczeństwo zwarcia, błędnej interpretacji zadania problemowego jako pisania kodu, który bezpośrednio rozwiązuje pewne konkretne wyzwanie postawione umysłowi, zamiast budowania umysłu, który może rozwiązać wyzwanie za pomocą ogólnej inteligencji. Kod, gdy jest nadużywany, jest doskonałym narzędziem do tworzenia długoterminowych problemów pod przykrywką krótkoterminowych rozwiązań. Opisując, czego nie wolno nam robić z kodem, jakie uzasadnione wyzwania leżą na tym poziomie organizacji? Niektóre wyzwania programistyczne są uniwersalne. Każdy współczesny programista powinien być zaznajomiony ze światem kompilatorów, interpreterów, debugerów, zintegrowanych środowisk programistycznych, programowania wielowątkowego, orientacji obiektowej, ponownego wykorzystania kodu, konserwacji kodu i innych narzędzi i tradycji współczesnego programowania. Trudno sobie wyobrazić, aby ktokolwiek z powodzeniem zakodował poziom brainware ogólnej inteligencji w języku asemblera – przynajmniej jeśli kod jest rozwijany po raz pierwszy. W tym sensie orientacja obiektowa i inne cechy współczesnych języków są „wymagane" do rozwoju AI; ale są niezbędne jako narzędzia produktywności, a nie ze względu na jakiekolwiek głębokie podobieństwo między strukturą języka programowania a strukturą ogólnej inteligencji. Dobre narzędzia programistyczne pomagają w rozwoju sztucznej inteligencji (AI), ale nie pomagają w samej sztucznej inteligencji.Niektóre wyzwania programistyczne, choć uniwersalne, prawdopodobnie będą niezwykle poważne w rozwoju AI. Rozwój AI jest eksploracyjny, paralelizowany i duży. Pisanie dużej ilości kodu eksploracyjnego oznacza, że IDE z obsługą refaktoryzacji i kontroli wersji są ważne, a kod modułowy jest jeszcze ważniejszy niż zwykle – lub przynajmniej kod, który jest tak modułowy, jak to możliwe, biorąc pod uwagę wysoce połączoną naturę supersystemu poznawczego. Paralelizm na poziomie sprzętowym jest obecnie obsługiwany przez symetryczne architektury układów wieloprocesorowych, klastrowanie NOW (sieć stacji roboczych) i klastrowanie Beowulf oraz interfejsy API do przekazywania wiadomości, takie jak PVM i MPI . Jednak współczesne języki nie radzą sobie dobrze z paralelizmem na poziomie oprogramowania i dlatego prawdopodobnie będzie stanowił jedno z największych wyzwań. Nawet gdyby paralelizm oprogramowania był dobrze obsługiwany, programiści AI nadal będą musieli poświęcić czas na wyraźne myślenie o tym, jak paralelizować procesy poznawcze – ludzkie poznanie może być masowo równoległe na niższych poziomach, ale ogólny przepływ poznania jest nadal szeregowy. Na koniec, istnieją pewne wyzwania programistyczne, które prawdopodobnie będą unikalne dla AI. Wiemy, że możliwe jest rozwinięcie ogólnej inteligencji, która działa na stu bilionach synaps z charakterystycznymi prędkościami granicznymi wynoszącymi około 200 impulsów na sekundę. Ciekawą właściwością ludzkiej neurobiologii jest to, że przy prędkości granicznej wynoszącej 150 metrów na sekundę dla zmielinizowanych aksonów, każdy neuron znajduje się potencjalnie w odległości mniej więcej jednego „tyknięcia zegara" od dowolnego innego neuronu w mózgu13. [90] opisuje wielkość S, która przekłada się na czas oczekiwania, w cyklach zegara, między różnymi częściami systemu poznawczego – minimalny czas, jaki zajmie sygnałowi przebycie drogi między najbardziej odległymi częściami systemu, mierzony w tyknięciach zegara systemu. W przypadku ludzkiego mózgu S jest mniej więcej rzędu 1 – przynajmniej w teorii. W praktyce aksony zajmują miejsce, a zmielinizowane aksony zajmują jeszcze więcej miejsca, więc mózg wykorzystuje wysoce modułową architekturę, ale nadal istnieją rury o dużym zasięgu, takie jak ciało modzelowate. Obecnie S jest znacznie większe niż 1 w przypadku klastrowanych systemów obliczeniowych. S jest większe niż 1 nawet w ramach systemu komputerowego z jednym procesorem; prawo Moore'a dotyczące przepustowości komunikacji wewnątrzsystemowej opisuje znacznie wolniejszy czas podwajania niż prędkość procesora. Coraz częściej ograniczającym zasobem nowoczesnych systemów obliczeniowych nie jest prędkość procesora, ale przepustowość pamięci (i ten problem pogorszył się, a nie poprawił, od 1995 r.). Jedna klasa czysto programowych problemów, które są unikalne dla AI, wynika z potrzeby „przenoszenia" inteligencji z masowo równoległych neuronów do klastrowanych systemów

obliczeniowych (lub innego programowalnego przez człowieka podłoża). Można sobie na przykład wyobrazić, że ludzki umysł obsługuje proces poznawczy skojarzenia pamięci poprzez porównywanie bieżących obrazów roboczych ze wszystkimi przechowywanymi pamięciami równolegle. Nie mamy żadnych konkretnych dowodów na to, że ludzki umysł używa porównania siłowego, ale może być ono siłowe. Ludzki mózg nie uznaje rozróżnienia między procesorem a pamięcią RAM. Jeśli jest wystarczająco dużo neuronów do przechowywania pamięci, to te same neurony mogą być prawdopodobnie wzywane do porównania tej pamięci z bieżącym doświadczeniem. (To prawda, nawet jeśli korespondencja między grupami neuronowymi a przechowywanymi wspomnieniami jest wiele do wielu, a nie jeden do jednego). Skojarzenie pamięci może lub nie używać operacji „porównania" (siłowego lub innego) bieżących obrazów ze wszystkimi przechowywanymi wspomnieniami, ale wydaje się prawdopodobne, że mózg używa masowo równoległego algorytmu w jednym lub drugim punkcie swojej operacji; skojarzenie pamięci jest po prostu prawdopodobnym kandydatem. Załóżmy, że skojarzenie pamięci jest zadaniem siłowym, wykonywanym przez poproszenie wszystkich neuronów zaangażowanych w przechowywanie pamięci o wykonanie „porównania" ze wzorcami nadawanymi z bieżących obrazów roboczych. Stając w obliczu wymogu projektowego dopasowania brutalnej siły 1014 masowo równoległych synaps do zwykłego systemu klastrowego, programista może być skłonny do rozpaczy. Nie ma żadnego a priori powodu, dla którego takie zadanie miałoby być możliwe. Stając w obliczu problemu tej klasy, programista może podjąć dwie drogi. Pierwsza to zaimplementowanie analogicznego „masowego porównania" tak wydajnie, jak to możliwe na dostępnym sprzęcie – wyzwanie algorytmiczne godne Herkulesa, ale poprzedni programiści pokonali ogromne bariery obliczeniowe dzięki bohaterskim wysiłkom i nieustannemu szlifowaniu prawa Moore'a. Druga droga – znacznie straszniejsza, z jeszcze mniejszą gwarancją, że sukces jest możliwy – to przeprojektowanie procesu poznawczego dla innego sprzętu. Najbardziej podstawowym ograniczeniem ludzkiego mózgu jest jego szybkość. Wszystko, co dzieje się w czasie krótszym niż sekunda, musi z konieczności wykorzystywać mniej niż 200 kolejnych operacji, niezależnie od tego, jak bardzo jest masowo sparalelizowane. Jeśli ludzki mózg naprawdę wykorzystuje masowo równoległe porównanie siłowe ze wszystkimi przechowywanymi pamięciami, aby poradzić sobie z problemem skojarzeń, to prawdopodobnie dlatego, że nie ma czasu na nic innego! Ludzki mózg jest masowo równoległy, ponieważ masowy paralelizm to jedyny sposób, aby zrobić cokolwiek w ciągu 200 taktów zegara. Gdyby współczesne komputery działały z częstotliwością 200 Hz zamiast 2 GHz, komputery PC potrzebowałyby również 1014 procesorów, aby robić cokolwiek interesującego w czasie rzeczywistym. Wystarczająco odważny ogólny programista sztucznej inteligencji, zamiast próbować ponownie zaimplementować poznawczy proces skojarzeń, który rozwinął się u ludzi, mógłby zamiast tego zapytać: Jak wyglądałby ten podsystem poznawczy, gdyby ewoluował na sprzęcie, a nie na oprogramowaniu? Jeśli usuniemy stare ograniczenie konieczności ukończenia w ciągu kilku tyknięć zegara i dodamy nowe ograniczenie niemożności „paralelizowania ze wszystkimi przechowywanymi pamięciami", jaki jest nowy najlepszy algorytm dla skojarzeń pamięci? Na przykład załóżmy, że znajdziesz metodę „rozmytego hashowania" pamięci, tak że większość podobnych pamięci automatycznie koliduje w przestrzeni kontenera, ale gdzie rozmyty hash z natury wymaga rozszerzonej liniowej serii kolejnych operacji, które umieściłyby „rozmyte hashowanie" poza zasięgiem dla operacji neuronowych w czasie rzeczywistym. „Rozmyte hashowanie" byłoby wówczas silnym kandydatem na alternatywną implementację skojarzeń pamięci. Tańszy obliczeniowo podsystem skojarzeń, który wykorzystuje prędkość szeregową zamiast prędkości równoległej, czy to oparty na „rozmytym hashowaniu", czy na czymś zupełnie innym, może być nadal jakościowo mniej inteligentny niż odpowiadający mu system skojarzeń w ludzkim mózgu. Na przykład rozpoznawanie pamięci może być ograniczone do kontekstów klastrowych, a nie być w pełni ogólne w odniesieniu do wszystkich przeszłych doświadczeń, przy czym SI często pomija „oczywiste" skojarzenia (gdzie „oczywiste" ma antropocentryczne znaczenie „obliczeniowo łatwe dla ludzkiego obserwatora"). W tym przypadku

pytanie brzmi, czy ogólna inteligencja mogłaby działać wystarczająco dobrze, aby sobie poradzić, być może rekompensując brak szerokości asocjacyjnej poprzez stosowanie dłuższych liniowych łańcuchów rozumowania. Różnica między serializmem a paralelizmem, na niskim poziomie, rozprzestrzeniałaby się w górę, tworząc różnice poznawcze, które rekompensują utratę ludzkich zalet lub wykorzystują nowe zalety, których ludzie nie dzielą. Inna klasa problemów wynika z „przenoszenia" przez skrajnie różne style programowania ewolucji w porównaniu z kodowaniem ludzkim. Programy pisane przez ludzi zazwyczaj obejmują długą serię powiązanych zależności, które przecinają się w pojedynczych punktach awarii – „krystaliczny" to dobry termin opisujący większość ludzkiego kodu. Obliczenia w neuronach mają inny charakter. Z biegiem czasu nasze obrazy neuronów biologicznych ewoluowały od prostych integratorów sygnałów synaptycznych, które uruchamiają się po osiągnięciu progowego poziomu sygnału wejściowego, do wyrafinowanych procesorów biologicznych z mieszaną logiką analogowo-cyfrową, adaptacyjną plastycznością, obliczeniami dendrytycznymi i funkcjonalnie istotnymi morfologiami dendrytycznymi i synaptycznymi [50]. Prawdą pozostaje, że z perspektywy algorytmicznej obliczenia neuronowe wykorzystują mniej więcej arytmetyczne operacje które przebiegają wzdłuż wielu przeplatających się kanałów, w których informacje są reprezentowane redundantnie i przetwarzane stochastycznie. Stąd łatwiej jest „trenować" sieci neuronowe – nawet niebiologiczne sieci koneksjonistyczne – niż trenować fragment kodu napisanego przez człowieka. Przewrócenie losowego bitu w stanie uruchomionego programu lub przewrócenie losowego bitu w instrukcji języka asemblera ma znacznie większy wpływ niż podobne zaburzenie sieci neuronowej. W przypadku sieci neuronowych krajobrazy sprawności są gładsze. Dlaczego tak jest? Biologiczne sieci neuronowe muszą tolerować większy szum środowiskowy (błąd danych) i szum procesora (błąd obliczeniowy), ale to dopiero początek wyjaśnień. Gładkie krajobrazy dostosowania są użytecznym, koniecznym i fundamentalnym wynikiem ewolucji. Każdy sukces ewolucyjny zaczyna się jako mutacja – błąd – lub jako nowa kombinacja genetyczna. Współczesny organizm, silnie adaptacyjny z dużym rezerwuarem złożoności genetycznej, koniecznie posiada bardzo długą historię ewolucyjną; to znaczy, że genotyp koniecznie przeszedł przez bardzo dużą liczbę udanych mutacji i rekombinacji na drodze do swojej obecnej formy. „Ewolucja ewolucyjności" jest najczęściej uzasadniana odniesieniem do tego historycznego ograniczenia [16], ale podejmowano również próby wykazania lokalnych nacisków selekcyjnych na cechy, które powodują ewolucyjność, unikając w ten sposób potrzeby powoływania się na kontrowersyjną agencję selekcji gatunków. Tak czy inaczej, gładkie krajobrazy dostosowania są częścią sygnatury projektowej ewolucji. „Płynne krajobrazy sprawności" oznaczają między innymi, że niewielkie zaburzenie w kodzie programu (szum genetyczny), w danych wejściowych (szum środowiskowy) lub w stanie wykonywanego programu (szum procesora) prawdopodobnie spowoduje co najwyżej niewielkie pogorszenie jakości wyjściowej. W większości kodów pisanych przez ludzi niewielkie zaburzenie dowolnego rodzaju zwykle powoduje awarię. Genomy są budowane przez kumulatywną serię mutacji punktowych i losowych rekombinacji. Programy pisane przez ludzi zaczynają się jako cele wysokiego poziomu, które są tłumaczone, przez rozszerzony szeregowy proces myślowy, na kod. Zaburzenie kodu pisanego przez ludzi zakłóca ostateczną formę kodu, a nie jego pierwszą przyczynę, a ostateczna forma kodu nie ma historii udanych mutacji. Myśli, które dały początek kodowi, prawdopodobnie mają metrykę płynnej sprawności, w tym sensie, że niewielkie zaburzenie stanu umysłu programisty prawdopodobnie spowoduje kod, który jest co najwyżej trochę gorszy, a być może trochę lepszy. Ludzkie myśli, które są pierwotnym źródłem kodu pisanego przez ludzi, są odporne; sam kod jest kruchy. Wymarzonym rozwiązaniem byłby język programowania, w którym kod pisany przez ludzi, odgórnie, w jakiś sposób miałby gładkie krajobrazy sprawności, które są charakterystyczne dla narastającej złożoności ewolucyjnej, ale to prawdopodobnie o wiele za wiele, aby wymagać tego od języka programowania. Różnica między ewolucją a projektowaniem jest głębsza niż różnica między stochastycznymi obwodami neuronowymi a kruchymi architekturami układów scalonych. Z drugiej strony, używanie kruchych bloków konstrukcyjnych nie może pomóc, więc

rozwiązanie na poziomie języka może rozwiązać przynajmniej część problemu. Znaczenie gładkich krajobrazów sprawności jest prawdziwe dla wszystkich poziomów organizacji. Koncepcje i myśli nie powinny się psuć w wyniku małych zmian. Poziom kodu jest wyodrębniany, ponieważ płynność na poziomie kodu reprezentuje inny rodzaj problemu niż płynność na wyższych poziomach. Na wyższych poziomach płynność jest produktem prawidłowo zaprojektowanych procesów poznawczych; wyuczona koncepcja będzie miała zastosowanie do chaotycznych nowych danych, ponieważ została wyabstrahowana z chaotycznej bazy doświadczalnej. Biorąc pod uwagę, że złożoność SI leżąca na poziomie koncepcji wymaga płynnych krajobrazów sprawności, właściwą strategią jest powielenie płynności na tym poziomie – przyjęcie jako wymagania projektowego wysokiego poziomu, że SI produkuje odporne na błędy koncepcje oderwane od chaotycznych baz doświadczalnych. Na poziomie kodu obwody neuronowe są gładkie i stochastyczne ze względu na naturę neuronów i naturę ewolucyjnego projektowania. Programy pisane przez ludzi są ostre i kruche („krystaliczne") ze względu na naturę nowoczesnych architektur chipów i naturę programowania przez ludzi. Rozróżnienie to prawdopodobnie nie zostanie zatarte przez wysiłek programisty lub nowe języki programowania. Długoterminowym rozwiązaniem może być SI z sensoryczną modalnością kodu (patrz sekcja 3), ale prawdopodobnie nie będzie to możliwe na wczesnych etapach. Podstawowy poziom kodu „rzeczy" ludzkiego mózgu ma wbudowane wsparcie dla płynnych krajobrazów sprawności, a podstawowy poziom kodu „rzeczy" pisanych przez ludzi programów komputerowych nie. Tam, gdzie procesy ludzkie polegają na tym, że obwody neuronowe są automatycznie odporne na błędy i możliwe do trenowania, potrzeba dodatkowej pracy programistycznej, aby „przenieść" ten proces poznawczy na nowy substrat, gdzie nie ma wbudowanego wsparcia. Ostateczne rozwiązanie kompromisowe może mieć tolerancję błędów jako jedną z wielu wyraźnych cech projektowych, a nie tolerancję błędów naturalnie wyłaniającą się z poziomu kodu. Istnieją inne ważne cechy, które są również obsługiwane przez biologiczne sieci neuronowe – które są „naturalne" dla substratu neuronowego. Cechy te prawdopodobnie obejmują:

• Optymalizacja pod kątem powtarzających się problemów

• Uzupełnianie częściowych wzorców

• Rozpoznawanie podobieństw (wykrywanie statycznego powtórzenia wzorca)

• Rozpoznawanie rekurencji (wykrywanie powtórzenia czasowego)

• Wykrywanie klastrowania, identyfikacja klastra i sortowanie do zidentyfikowanych klastrów

• Szkolenie w zakresie rozpoznawania wzorców i uzupełniania wzorców

• Masowy paralelizm

Ponownie, nie oznacza to nie do pobicia przewagi dla biologicznych sieci neuronowych. W niektórych przypadkach wetware ma bardzo słabe wsparcie funkcji w porównaniu do współczesnego sprzętu. Współczesny sprzęt ma lepsze wsparcie dla

• Odbicie i ślady wykonania

• Bezstratna serializacja (przechowywanie i pobieranie) i bezstratne transformacje wzorców

• Bardzo precyzyjne obliczenia ilościowe

• Algorytmy niskiego poziomu, które obejmują rozszerzoną iterację, głęboką rekurencję i złożone rozgałęzienia

• „Ogromny serializm"; możliwość wykonywania setek milionów kolejnych kroków na sekundę

Wyzwaniem jest wykorzystanie nowych zalet w celu zrekompensowania utraty starych zalet i zastąpienie wsparcia na poziomie podłoża wsparciem na poziomie projektu. To kończy opis wyjątkowych problemów, które pojawiają się na poziomie kodu. Wyliczenie wszystkich problemów, które pojawiają się na poziomie kodu – na przykład serializacja bieżącej zawartości modalności sensorycznej w celu wydajnej transmisji do zduplikowanej modalności na innym węźle rozproszonej sieci – stanowiłoby co najmniej jedną trzecią kompletnego konstruktywnego opisu ogólnej SI. Ale programowanie nie jest całą pracą SI, być może nawet nie większością pracy SI; większość wysiłku potrzebnego do skonstruowania inteligencji zostanie poświęcona na popychanie SI do tworzenia pewnych koncepcji, przechodzenia pewnych doświadczeń, odkrywania pewnych przekonań i uczenia się różnych umiejętności wysokiego poziomu. Zadań tych nie da się wykonać za pomocą IDE. Zakodowanie niewłaściwej rzeczy z powodzeniem może bardziej zepsuć projekt SI niż jakakolwiek liczba błędów programistycznych. Uważam, że najważniejszą umiejętnością, jaką może posiadać programista SI, jest wiedza o tym, czego nie programować.

Poziom modalności

Ewolucyjny projekt modalności u ludzi

Większość studentów AI zna wysokopoziomowe procesy obliczeniowe co najmniej jednej ludzkiej modalności sensorycznej, wzroku, przynajmniej w takim stopniu, że znają „świat 2 1/2D" Davida Marra i koncepcję ekstrakcji warstwowych cech [65]. Dalsze badania w zakresie obliczeniowej neuronauki potwierdziły teorię Marra i uczyniły ją znacznie bardziej złożoną. Chociaż wielu autorów, w tym ja, używało określenia „kora wzrokowa", mówiąc o całej modalności wzrokowej, jest to jak mówienie o Stanach Zjednoczonych przez odniesienie do Nowego Jorku. Około 50% neokory u naczelnych innych niż człowiek jest poświęcone wyłącznie przetwarzaniu wzrokowemu, przy czym u makaka zidentyfikowano ponad 30 odrębnych obszarów wzrokowych . Główny strumień wzrokowy to strumień siatkówkowo-kolankowo-korowy, który biegnie od siatkówki do jądra kolankowatego bocznego, do kory prążkowanej i do wyższych obszarów wzrokowych. Poza korą wzrokową przetwarzanie dzieli się na dwa główne strumienie wtórne; strumień brzuszny kierujący się w stronę płata skroniowego w celu rozpoznania obiektów i strumień grzbietowy kierujący się w stronę płata ciemieniowego w celu przetwarzania przestrzennego. Strumień wzrokowy zaczyna się w siatkówce, która zawiera około 100 milionów pręcików i 5 milionów czopków, ale zasila kabel optyczny zawierający tylko około 1 miliona aksonów. Wstępne przetwarzanie wizualne zaczyna się w pierwszej warstwie siatkówki, która przekształca surowe natężenia w gradienty centrum-otoczenia, reprezentację, która stanowi podstawę wszelkiego dalszego przetwarzania wizualnego. Po kilku kolejnych warstwach przetwarzania siatkówkowego, końcowa warstwa siatkówki składa się z szerokiej gamy typów zwojów, które obejmują kierunkowo selektywne detektory ruchu, wolno poruszające się detektory krawędzi, detektory szybkiego ruchu, detektory jednorodności i subtraktywne kanały kolorów. Aksony tych zwojów tworzą nerw wzrokowy i rzutują do warstw wielkokomórkowej, parwokomórkowej i koniokomórkowej jądra kolankowatego bocznego; obecnie wydaje się, że każda klasa zwojów rzutuje tylko do jednej z tych warstw. Powszechnie przyjmuje się, że dalsze wykrywanie cech ma miejsce w jądrze kolankowatym bocznym, ale szczegóły nie są obecnie jasne. Z jądra kolankowatego bocznego strumień informacji wzrokowych przechodzi do obszaru V1, pierwotnej kory wzrokowej, która rozpoczyna ekstrakcję cech dla informacji o ruchu, orientacji, kolorze i głębi. Z pierwotnej kory wzrokowej strumień informacji przechodzi dalej, kierując się do wyższych obszarów wzrokowych, od V2 do V6. Poza korą wzrokową strumień informacji przechodzi do obszarów skroniowych (rozpoznawanie obiektów) i obszarów ciemieniowych (przetwarzanie przestrzenne). Jak wspomniano wcześniej, pierwotna kora wzrokowa wysyła masywne projekcje sprzężenia zwrotnego korowo-wzgórzowego do jądra kolankowatego bocznego [94]. Połączenia korowo-korowe są również zwykle

połączone z projekcjami sprzężenia zwrotnego o równej sile. Obecnie nie ma standardowego wyjaśnienia tych połączeń sprzężenia zwrotnego. DGI16 wymaga modalności sensorycznych z kontrolerami cech, które są odwrotnymi uzupełnieniami detektorów cech; pasuje to do istnienia projekcji sprzężenia zwrotnego. Należy jednak zauważyć, że to stwierdzenie nie jest częścią współczesnej neuronauki. Istnienie kontrolerów cech jest dozwolone, ale nie stwierdzone przez obecną teorię; ich istnienie jest stwierdzone i wymagane przez DGI. (Hipoteza, że projekcje sprzężenia zwrotnego odgrywają rolę w obrazowaniu umysłowym, nie ogranicza się do DGI; na przykład [53] cytuje istnienie projekcji sprzężenia zwrotnego korowo-korowego jako dostarczających podstawowego mechanizmu dla funkcji poznawczych wyższego poziomu w celu kontrolowania obrazowych obrazów umysłowych). Ogólna lekcja wyciągnięta z ludzkiej modalności wzrokowej jest taka, że modalności nie są mikroteoriami, że modalności nie są płaskimi reprezentacjami poziomu pikseli i że modalności są funkcjonalnie charakteryzowane przez kolejne warstwy coraz bardziej złożonej struktury cech. Modalności są jedną z najlepszych wystaw tego ewolucyjnego wzorca projektowego – wstępujące warstwy adaptacyjnej złożoności – która pojawia się również, choć w bardzo odmiennej formie, w wstępującym modelu kod-modalność-koncept-myśl-rozważanie ludzkiego umysłu. Każda wstępująca warstwa jest bardziej rozbudowana, bardziej złożona, bardziej elastyczna i bardziej kosztowna obliczeniowo. Każda warstwa wymaga złożoności warstwy poniżej – zarówno funkcjonalnie w obrębie pojedynczego organizmu, jak i ewolucyjnie w obrębie populacji genetycznej. Warstwa koncepcyjna jest możliwa do rozwinięcia w serii krótkich kroków, jeśli i tylko jeśli istnieje już znaczna złożoność w obrębie warstwy modalności. Ten sam wzór projektowy – rosnące warstwy adaptacyjnej złożoności – pojawia się również w obrębie rozwiniętej modalności sensorycznej. Pierwsze wykryte cechy są proste i mogą ewoluować w jednym kroku lub w małej serii adaptacyjnych krótkich kroków. Możliwość wykrycia tych pierwszych cech może być adaptacyjna nawet w przypadku braku kompletnej modalności sensorycznej. Oko, o którym obecnie uważa się, że ewoluowało niezależnie u wielu różnych gatunków, mogło zaczynać za każdym razem jako pojedynczy, wrażliwy na światło punkt na skórze organizmu. W modalnościach każda dodatkowa warstwa detektorów cech wykorzystuje informacje dostarczane przez pierwszą warstwę detektorów cech. W przypadku braku pierwszej warstwy detektorów cech „kod" dla drugiej warstwy detektorów cech byłby zbyt złożony, aby ewoluować w jednym fragmencie. Gdy pierwsza warstwa detektorów cech jest już obecna, detektory cech w drugiej warstwie mogą ewoluować w jednym kroku lub w krótkiej serii lokalnie adaptacyjnych kroków. Kolejne warstwy organizacji w modalności sensorycznej są piękną ilustracją sygnatury projektu ewolucji, funkcjonalnej ontogenezy informacji podsumowującej ewolucyjną filogenezę. Ewolucja jest dobrym nauczycielem, ale kiepskim wzorem do naśladowania; czy ten projekt jest błędem czy cechą? Twierdziłbym, że ogólnie jest cechą. Istnieje głęboka korespondencja między ewolucyjnie gładkimi krajobrazami sprawności a obliczeniowo gładkimi krajobrazami sprawności. Istnieje głęboka korespondencja między każdą kolejną warstwą detektorów cech, która jest ewolucyjna, a każdą kolejną warstwą detektorów cech, która jest obliczalna w sposób, który jest „gładki", a nie „kruchy", jak opisano we wcześniejszej dyskusji na temat warstwy kodu. Gładkie obliczenia są bardziej ewolucyjne, więc ewolucja, konstruując system stopniowo, ma tendencję do konstruowania liniowych sekwencji lub rosnących warstw gładkich operacji. Projektant AI może teoretycznie odrzucić wymóg, aby każda wstępująca warstwa wykrywania cech była stopniowo użyteczna/adaptacyjna – chociaż może to utrudnić stopniowe rozwijanie i testowanie podsystemu! Jednak poznawczo ważne jest, aby kolejne warstwy detektorów cech były obliczeniowo „gładkie" w jednym konkretnym sensie. Koncepcje DGI oddziałują z odwrotnymi detektorami cech, „kontrolerami cech", w celu konstruowania obrazów mentalnych. Aby zadanie narzucania koncepcji i jeszcze trudniejsze zadanie abstrahowania koncepcji były jednocześnie wykonalne, konieczne jest, aby modalności sensoryczne były kontinuum lokalnie gładkich warstw, a nie składały się z ogromnych, nieuchwytnych, nieprzezroczystych fragmentów. Istnieje głęboka korespondencja między gładkim

projektem, który sprawia, że koncepcje są wykonalne, a gładką architekturą wyłaniającą się z przyrostowej ewolucji. Kontrolery cech używane do tworzenia obrazów mentalnych są ewolucyjne i preadaptacyjne w przypadku braku obrazów mentalnych; kontrolery cech mogłyby zaczynać się jako ograniczenia odgórne w przetwarzaniu percepcyjnym lub jeszcze prościej jako krok percepcyjny, który jest najlepiej obliczany przez sieć rekurencyjną. W obu przypadkach najłatwiejsza (najbardziej ewoluująca) architektura to na ogół taka, w której połączenie sprzężenia zwrotnego odwzajemnia połączenie sprzężenia zwrotnego. Tak więc warstwy kontrolerów cech nie są oddzielnym systemem niezależnym od warstw detektorów cech; raczej spodziewam się, że to, co jest lokalnie detektorem cech, jest również lokalnie kontrolerem cech. Ponownie, ta płynna odwracalność pomaga uczynić możliwym nauczenie się pojedynczego pojęcia, które może działać jako detektor kategorii lub narzucający kategorię. To jednoczesne rozwiązanie narzucania pojęć, zadowolenia pojęć, fasetowania pojęć i abstrakcji pojęć wymaga odwracalnych cech – kontrolerów cech, które są lokalnymi odwrotnościami detektorów cech. Wątpię, aby kontrolery cech docierały aż do pierwszych warstw siatkówki (nie słyszałem o żadnych połączeniach sprzężenia zwrotnego sięgających tak daleko), ale bezpośrednie dowody z neuroobrazowania pokazują, że obrazowanie umysłowe aktywuje pierwotną korę wzrokową; Nie jestem pewien, czy analogiczne testy przeprowadzono dla jądra kolankowatego bocznego, ale połączenia sprzężenia zwrotnego tam występują.

Ludzki projekt modalności w AI

AI potrzebuje modalności sensorycznych – ale jakich modalności? W jaki sposób te modalności przyczyniają się materialnie do ogólnej inteligencji poza bezpośrednią modalnością? Czy AI potrzebuje systemu wizualno-przestrzennego wzorowanego na wielkiej złożoności układu wizualno-przestrzennego u naczelnych i ludzi? Wiemy więcej o ludzkiej modalności wzrokowej niż o jakimkolwiek innym aspekcie neurologii człowieka, ale nie oznacza to, że wiemy wystarczająco dużo, aby zbudować modalność wizualną od podstaw. Co więcej, ludzka modalność wzrokowa jest ogromnie złożona, intensywnie obliczeniowo i dostosowana do środowiska, którego AI niekoniecznie musi natychmiast zrozumieć. Czy ludzki, trójwymiarowy wzrok powinien być jedną z pierwszych próbowanych modalności? Uważam, że najlepiej będzie odrzucić ludzkie modalności lub wykorzystać je wyłącznie jako inspirację – użyć zupełnie innego zestawu modalności sensorycznych na wczesnych etapach rozwoju AI. AI zajmuje inne środowisko niż człowiek, a bezpośrednie naśladowanie ludzkich modalności nie byłoby właściwe. W przypadku początkowych doświadczeń uczenia się AI opowiadałbym się za umieszczeniem AI w złożonych środowiskach wirtualnych, w których środowiska wirtualne są wewnętrzne dla komputera, ale zewnętrzne dla AI. Następnie programiści próbowaliby opracować modalności sensoryczne odpowiadające środowiskom wirtualnym. Odtąd mogę używać terminu „mikrośrodowisko", aby wskazać złożone środowisko wirtualne. Termin „mikroświat" jest mniej nieporęczny, ale nie należy go rozumieć jako mającego konotację „mikroświatów" w dobrej, staromodnej AI, w której wszystkie cechy są bezpośrednio reprezentowane przez logikę predykatów, np. uproszczony świat bloków i tabel SHRDLU. Porzucenie modalności ludzkich wydaje się wprowadzać dodatkową kruchą zależność od poprawności teorii AI, w tym sensie, że zastąpienie modalności ludzkich nowymi modalnościami sensorycznymi wydaje się wymagać prawidłowego zrozumienia natury modalności sensorycznych i tego, w jaki sposób przyczyniają się one do inteligencji. To prawda, ale argumentowałbym, że istnienie dodatkowej zależności jest iluzoryczne. Próba ślepego naśladowania ludzkiej modalności wzrokowej, bez zrozumienia roli modalności w inteligencji, raczej nie przyczyniłaby się do ogólnej inteligencji, chyba że przez przypadek. Nasze współczesne rozumienie ludzkiej modalności wzrokowej nie jest tak doskonałe, abyśmy mogli polegać na funkcjonalnej kompletności projektu inspirowanego neurologicznie; na przykład projekt oparty wyłącznie na konsensusie współczesnej teorii mógłby pominąć kontrolery cech! Jednak przejście do mikroświatów wymaga, aby doświadczenie w mikroświatach odtwarzało funkcjonalnie istotne aspekty doświadczenia

w prawdziwym życiu, w tym nieprzewidywalność, niepewność, kontrolę procesu w czasie rzeczywistym, organizację holoniczną (część-całość) itd. Nie sądzę, aby wprowadzało to dodatkową zależność od teoretycznego zrozumienia, wykraczającą poza teoretyczne zrozumienie, które byłoby wymagane do zbudowania SI, która absorbowałaby złożoność z tych aspektów środowisk świata rzeczywistego, ale mimo to stanowi silną zależność od teoretycznego zrozumienia. Załóżmy, że projektujemy de novo modalność sensoryczną i środowisko wirtualne. Trzy możliwe modalności, które przychodzą na myśl jako rozsądne dla bardzo prymitywnej i wczesnej fazy AI, w kolejności rosnącej trudności implementacji, to:

1. Modalność dla kul bilardowych Newtona

2. Modalność dla planszy „Go" 100x100

3. Modalność dla pewnego rodzaju interpretowanego kodu (metaforyczna „kora kodowa")

W ludzkim wzroku pierwszymi neuronami wzrokowymi są „pręciki i czopki", które przetwarzają uderzające fotony środowiskowe na reprezentację neuronalną jako informacje sensoryczne. W przypadku każdej z trzech powyższych modalności poziom „pręcików i czopków" prawdopodobnie używałby zasadniczo tej samej reprezentacji, co struktury danych używane do tworzenia mikroświata lub środowiska wirtualnego, w którym ucieleśniona jest sztuczna inteligencja. Jest to duże odejście od projektu naturalnie ewoluujących modalności, w których podstawowy poziom – poziom kwarków, o ile nam wiadomo – to wiele warstw usuniętych z obiektów wysokiego poziomu, które dają początek pośrednim informacjom docierającym do zmysłów. Ewoluujące modalności sensoryczne poświęcają większość swojej złożoności na rekonstrukcję świata, który daje początek przychodzącym wrażeniom sensorycznym – na rekonstrukcję trójwymiarowych ruchomych obiektów, które dają początek fotonom uderzającym w warstwę pręcików i czopków siatkówki. Oczywiście wybór wzroku jako przykładu jest prawdopodobnie stronniczym wyborem; dźwięk nie jest tak złożony jak wzrok, a węch i smak nie są tak złożone jak dźwięk. Niemniej jednak wyeliminowanie niepewności i pośrednich warstw pomiędzy prawdziwym środowiskiem a danymi sensorycznymi organizmu jest ważnym krokiem. Powinno to znacznie zmniejszyć wyzwania wczesnego rozwoju SI, ale jest to jednak niebezpieczny krok ze względu na jego odległość od paradygmatu biologicznego i eliminację znaczącego źródła złożoności. Zalecam wyeliminowanie rekonstrukcji środowiska jako źródła złożoności we wczesnym rozwoju SI. Wizualizacja perspektywy celowego pogorszenia jakości informacji środowiskowych SI z jednej strony i rozwinięcia modalności sensorycznej SI z drugiej strony, uważam za prawdopodobne, że cała operacja zostanie anulowana, nie wnosząc niczego. SI, która musiałaby nauczyć się rekonstruować środowisko w taki sam sposób, w jaki ewolucja nauczyła się konstruować modalności sensoryczne, mogłaby w rezultacie wytworzyć interesującą złożoność; ale jeśli ten sam programista tworzy złożoność środowiskową i złożoność modalności, spodziewałbym się, że te dwie operacje zostaną anulowane. Podczas gdy rekonstrukcja środowiska jest nietrywialnym źródłem złożoności w ludzkim mózgu, uważam, że stosunek trudności programistycznego rozwoju złożoności do wkładu tej złożoności w ogólną inteligencję jest stosunkowo niewielki. Dodanie złożoności do rekonstrukcji środowiska poprzez wprowadzenie dodatkowych warstw złożoności w mikroświecie i celowe wprowadzenie strat informacji między najwyższą warstwą mikroświata a receptorami sensorycznymi SI, a następnie próba stworzenia modalności SI, która mogłaby zrekonstruować oryginalną treść mikroświata z końcowego sygnału sensorycznego, wymagałoby stosunkowo dużego nakładu wysiłku w zamian za to, co podejrzewam, byłoby stosunkowo niewielkim zwiększeniem ogólnej inteligencji. Załóżmy, że dla każdej z trzech modalności – bilard, Go, kod – poziom „przedsiatkówkowy" składa się z prawdziwych i dokładnych informacji o poziomie kwarków wirtualnego mikroświata, choć być może nie są to kompletne informacje, a zasadnicza złożoność, która czyni model „modalnością sensoryczną", spoczywa w strukturze cech, wstępujących warstwach detektorów cech i zstępujących warstwach

kontrolerów cech. Które cechy są zatem odpowiednie? A w jaki sposób przyczyniają się one w sposób istotny do ogólnej inteligencji? Zwykłym stwierdzeniem jest, że złożoność w modalności sensorycznej odzwierciedla regularności otoczenia, ale chciałbym przedstawić nieco inny punkt widzenia. Aby zilustrować ten pogląd, muszę pożyczyć i znacznie uprościć puentę naprawdę eleganckiego artykułu „The Perceptual Organization of Colors" autorstwa Rogera Sheparda . Oprócz innych pytań, artykuł ten stara się odpowiedzieć na pytanie o trichromancję: Dlaczego w siatkówce człowieka znajdują się trzy rodzaje czopków, a nie dwa lub cztery? Dlaczego ludzkie postrzeganie wzrokowe jest zorganizowane w trójwymiarową przestrzeń kolorów? Historycznie często teoretyzowano, że trichromancja stanowi arbitralny kompromis między rozdzielczością chromatyczną a rozdzielczością przestrzenną; to znaczy między liczbą postrzeganych kolorów a wielkością ziarna rozdzielczości wizualnej. Jak się okazuje, istnieje bardziej fundamentalny powód, dla którego potrzebne są trzy kanały kolorów. Aby wyjaśnić pytanie, rozważmy, że powierzchnie posiadają potencjalnie nieskończoną liczbę rozkładów odbicia widmowego. Skupimy się na rozkładach odbicia widmowego, a nie na rozkładach mocy widmowej, ponieważ obiekty adaptacyjnie istotne, które emitują własne światło, są rzadkie w środowisku. Stąd fizycznie stałą właściwością większości obiektów jest rozkład odbicia widmowego, który łączy się z rozkładem mocy widmowej światła padającego na obiekt, dając początek rozkładowi mocy widmowej odbieranemu przez ludzkie oko. Rozkład odbicia widmowego jest definiowany w zakresie długości fal od 400 nm do 700 nm (zakres widzialny), a ponieważ długość fali jest ciągła, rozkład odbicia widmowego może teoretycznie wymagać nieograniczonej liczby wielkości do określenia. Stąd nie jest możliwe dokładne ograniczenie rozkładu odbicia widmowego przy użyciu tylko trzech wielkości, czyli ilości informacji przetwarzanych przez ludzkie czopki. Ludzkie oko nie jest w stanie rozróżnić wszystkich fizycznie możliwych powierzchni odbijających. Jednakże możliwe jest, że w przypadku „naturalnych" powierzchni – powierzchni powszechnie spotykanych w środowisku przodków – odbicie dla każdej czystej częstotliwości nie zmienia się niezależnie od odbicia dla wszystkich innych częstotliwości. Na przykład może istnieć pewien zestaw bazowych funkcji odbicia, tak że rozkłady odbicia prawie wszystkich naturalnych powierzchni można wyrazić jako ważoną sumę wektorów bazowych. Jeśli tak, jednym z możliwych wyjaśnień trichromacji ludzkiego wzroku byłoby to, że trzy kanały kolorów wystarczają do przeprowadzenia odpowiedniego rozróżnienia w „naturalnej" przestrzeni kolorów o ograniczonej wymiarowości. Zdolność do rozróżniania wszystkich naturalnych powierzchni byłaby projektem zalecanym przez filozofię „regularności środowiskowej" modalności sensorycznych. Wymiarowość wewnętrznego modelu odzwierciedlałaby wymiarowość środowiska. Okazuje się, że naturalne powierzchnie mają widmowe rozkłady odbicia, które zmieniają się w przybliżeniu w pięciu do siedmiu wymiarach. Istnieją zatem naturalne powierzchnie, które, chociaż wydają się trójchromatycznym obserwatorom „tego samego koloru", mimo to posiadają różne widmowe rozkłady odbicia. Pada pytanie, ile kanałów kolorów jest potrzebnych, aby zapewnić, że kolor, który postrzegamy, jest tym samym kolorem za każdym razem, gdy powierzchnia jest oglądana w różnych warunkach oświetleniowych. Ilość światła otoczenia może również potencjalnie zmieniać się wzdłuż nieograniczonej liczby wymiarów, a rzeczywiste światło docierające do oka jest iloczynem rozkładu mocy widmowej i rozkładu odbicia widmowego. Czerwonawy obiekt w niebieskim świetle może odbijać taką samą liczbę fotonów każdej długości fali, jak niebieski obiekt w czerwonym świetle. Podobnie, biały obiekt w czerwonym świetle może odbijać głównie czerwone fotony, podczas gdy ten sam biały obiekt w niebieskim świetle może odbijać głównie niebieskie fotony. A jednak ludzki układ wzrokowy potrafi utrzymać właściwość stałości koloru; ten sam obiekt będzie wydawał się mieć ten sam kolor w różnych warunkach oświetleniowych. Zmierzono 622 rozkłady mocy widmowej dla naturalnego oświetlenia, w 622 bardzo zmiennych naturalnych warunkach pogodowych i porach dnia, i stwierdził, że zmiany w naturalnym oświetleniu zmniejszają się do trzech stopni swobody. Co więcej, te trzy stopnie swobody ściśle odpowiadają trzem wymiarom przeciwstawności kolorów, które

zaproponowano dla ludzkiego układu wzrokowego na podstawie badań eksperymentalnych. Te trzy stopnie swobody to:

1. Zmiana światła i ciemności, która zależy od całkowitego światła docierającego do obiektu.

2. Zmiana żółto-niebieska, która zależy od tego, czy powierzchnia jest oświetlona bezpośrednim światłem słonecznym, czy znajduje się w cieniu. W cieniu powierzchnia jest oświetlona rozproszonym przez Raleigh niebieskim światłem nieba, ale nie jest bezpośrednio oświetlona przez słońce. Odpowiednia żółta skrajność występuje, gdy obiekt jest oświetlony tylko bezpośrednim światłem słonecznym; np. jeśli światło słoneczne wchodzi przez mały kanał, a światło nieba jest odcięte.

3. Zmiana czerwono-zielona, która zależy zarówno od wysokości słońca (ilości atmosfery, przez którą przechodzi słońce), jak i ilości pary wodnej w atmosferze. Np. oświetlenie przez czerwony zachód słońca w porównaniu z oświetleniem w południe. Czerwone długości fal to długości fal najmniej rozpraszane przez pył i najbardziej pochłaniane przez wodę.

Trzy kanały kolorów ludzkiego układu wzrokowego to dokładnie taka liczba kanałów, jaka jest potrzebna do utrzymania stałości kolorów w warunkach naturalnego oświetlenia. Trzy kanały kolorów nie wystarczą, aby odróżnić wszystkie naturalne odbicia powierzchni, ale trzy kanały kolorów to dokładna liczba wymagana do skompensowania naturalnego oświetlenia otoczenia i tym samym zapewnienia, że ta sama powierzchnia jest percepcyjnie „tego samego koloru" w dowolnych dwóch przypadkach. Upraszcza to adaptacyjnie ważne zadanie rozpoznawania wcześniej doświadczonego obiektu podczas przyszłych spotkań. Lekcja, jaką wyniosę z tej opowieści moralnej o stałości kolorów, to to, że modalności sensoryczne dotyczą niezmienników, a nie tylko regularności. Rozważmy zadanie zaprojektowania modalności sensorycznej dla jakiejś formy interpretowanego kodu. (Jest to bardzo trudne zadanie, ponieważ ludzkie języki programowania mają tendencję do niepłynnych krajobrazów dopasowania, jak wcześniej omówiono). Rozważając, które cechy wyodrębnić, pytanie, które bym zadał, nie brzmi: „Jakie regularności występują w kodzie?" ale raczej „Jaka struktura cech jest potrzebna, aby AI postrzegała dwa identyczne algorytmy z nieznacznie różniącymi się implementacjami jako „ten sam fragment kodu"?" Albo bardziej konkretnie: „Jakie cechy ta modalność musi wyodrębnić, aby postrzegać rekurencyjny algorytm dla ciągu Fibonacciego i iteracyjny algorytm dla ciągu Fibonacciego jako „ten sam fragment kodu"?" Lekko przechyl głowę w lewo, a następnie w prawo. Każdy receptor siatkówki może odbierać inny sygnał, ale doświadczane pole widzenia pozostaje niemal dokładnie takie samo". Podnieś pionek szachowy i lekko przechyl go w lewo lub w prawo. Pomimo zmian w odbiorze siatkówki widzimy „tego samego" pionka z nieznacznie inną orientacją. Czy modalność sensoryczna kodu mogłaby spojrzeć na dwa zestawy zinterpretowanych bajtkodów (lub innych list programów), całkowicie różnych bajt po bajcie, i zobaczyć te dwa listy jako „ten sam" algorytm w dwóch nieznacznie różnych „orientacjach"? Poziom organizacji modalności, podobnie jak poziom kodu, ma charakterystyczny rodzaj pracy, którą wykonuje. Sformułowanie koncepcji motyla i postrzeganie dwóch motyli jako członków tej samej kategorii jest pracą poziomu koncepcji, ale postrzeganie pionka szachowego w dwóch orientacjach jako tego samego pionka jest pracą poziomu modalności. Istnieje nakładanie się między poziomem modalności a poziomem koncepcji, tak jak istnieje nakładanie się między poziomem kodu a poziomem modalności. Jednak ogólnie rzecz biorąc, poziom modalności dotyczy raczej niezmienników niż regularności i tożsamości niż kategorii. Podobnie, rozumienia udzielanego przez poziom modalności nie należy mylić z analitycznym rozumieniem charakterystycznym dla myśli i rozważań. Wracając do przykładu modalności kodowej, jednym z możliwych wskaźników poważnego błędu projektowego byłoby skonstruowanie modalności, która mogłaby równie dobrze analizować każdy możliwy fragment kodu. Pierwsza warstwa siatkówki –

pręciki i czopki – jest jedyną częścią ludzkiego układu wzrokowego, która będzie działać na wszystkich możliwych polach pikseli. Reszta układu wzrokowego będzie działać tylko dla pól pikseli o niskiej entropii doświadczanych przez organizm o niskiej entropii w środowisku o niskiej entropii. Kolejna warstwa, po pręcikach i czopkach, opiera się już na organizacji centrum-otoczenie jako użytecznym sposobie kompresji informacji wzrokowych; jest to prawdą tylko w środowisku wzrokowym o niskiej entropii. Zaprojektowanie modalności, która działałaby równie dobrze dla każdego możliwego programu komputerowego, prawdopodobnie byłoby wskazówką, że modalność ta wydobywała niewłaściwy rodzaj informacji. Dlatego należy zachować ostrożność w przypadku rzekomej „struktury cech", która wygląda, jakby działała równie dobrze dla wszystkich możliwych fragmentów kodu. Może to być prawidłowa metoda analityczna, ale prawdopodobnie należy do poziomu rozważań, a nie poziomu modalności. (Trzeba przyznać, że nie każdy lokalny krok modalności musi być zależny od danych wejściowych o niskiej entropii; niektóre lokalne etapy przetwarzania mogą mieć matematyczną naturę bezstratnej transformacji, która działa równie dobrze na każdym możliwym danych wejściowych. Ponadto sprzęt może być lepiej przystosowany do bezstratnych transformacji niż oprogramowanie.)

Ludzki mózg jest ograniczony przez charakterystyczną prędkość szeregową 200 kolejnych kroków na sekundę i przez wszechobecne wewnętrzne wykorzystanie synchronicznego nadejścia skojarzonych informacji, aby zorganizować etapy przetwarzania, które płynnie przechodzą do przodu. Logika „jeśli-wtedy" lub „przypadku przełączania" wyższego poziomu jest trudniejsza do osiągnięcia neuronalnie, a rozszerzona złożona logika „jeśli-wtedy" lub „przypadku przełączania" jest prawdopodobnie prawie niemożliwa, chyba że zostanie zaimplementowana za pomocą rozgałęzionych obwodów równoległych, które pozostają zsynchronizowane. Prawdopodobnie wyjątkowy warunek musi zostać zignorowany, uśredniony lub w inny sposób obsłużony przy użyciu tych samych algorytmów, które miałyby zastosowanie do dowolnej innej treści modalności. Czy modalność AI może wykorzystywać architekturę, która stosuje różne algorytmy do różnych elementów treści modalności?

Czy modalność AI może obsługiwać wyjątkowe warunki za pomocą kodu specjalnego? Radziłbym zachować ostrożność z kilku powodów. Po pierwsze, główne gałęzie „jeśli-to" są charakterystyczne dla procesów deliberatywnych, a pokusa użycia takiej gałęzi może wskazywać na pewien poziom zamieszania. Po drugie, tworzenie wyjątków od płynnego przepływu przetwarzania prawdopodobnie skomplikuje łączenie pojęć i modalności. Po trzecie, modalności są niedoskonałymi, ale odpornymi na błędy procesami, a tolerancja błędów odgrywa rolę w wygładzaniu krajobrazów sprawności i umożliwianiu budowania wyższych poziomów organizacji na górze; zatem próba obsługi wszystkich danych poprzez wykrywanie wyjątkowych warunków i korygowanie ich, standardowy wzorzec w programowaniu ludzkim, może wskazywać, że modalność jest niewystarczająco odporna na błędy. Po czwarte, obsługa wszystkich wyjątków jest charakterystyczna dla próby obsługi wszystkich danych wejściowych, a nie tylko danych wejściowych o niskiej entropii. Stąd, ogólnie rzecz biorąc, modalności sensoryczne charakteryzują się płynnym przepływem informacji przez rosnące warstwy detektorów cech. Oczywiście wykrywanie wyjątkowego warunku jako cechy może okazać się całkowicie właściwe! Innym problemem, który może pojawić się w sztucznych modalnościach sensorycznych, jest to, że nieskomplikowane sztuczne modalności mogą okazać się znacznie droższe obliczeniowo w stosunku do efektywnej inteligencji, którą dostarczają. Skomplikowane ewolucyjne modalności oszczędzają moc obliczeniową w sposób, który może być bardzo trudny do powtórzenia przez ludzkiego programistę. Przykładem może być wykorzystanie częściowego obrazowania, modelowanie tylko cech, które są potrzebne do zadania wysokiego poziomu; uproszczona modalność, która nie obsługuje częściowego obrazowania, może zużywać więcej mocy obliczeniowej. Innym przykładem może być selektywna koncentracja ludzkiego układu wzrokowego na środku pola widzenia – „architektura dołka", w której obszary pola widzenia bliższe środka są przydzielane większej liczbie neuronów. Współczynnik

powiększenia korowego u naczelnych jest odwrotnie liniowy; zespolony logarytm jest jedyną dwuwymiarową funkcją mapy, która ma tę właściwość, co zostało potwierdzone eksperymentalnie przez . Stała rozdzielczość wersji kory wzrokowej, z maksymalną ludzką rozdzielczością wizualną w całym polu widzenia człowieka, wymagałaby 10 000 razy więcej komórek niż nasza rzeczywista kora [86]. Ale rozważmy problemy programowe wprowadzone przez użycie mapy logarytmicznej. W zależności od tego, gdzie obiekt znajduje się w polu widzenia, jego wewnętrzna reprezentacja na mapie retinotopowej będzie zupełnie inna; żadne bezpośrednie porównanie struktur danych nie pokazałoby tożsamości ani nawet nie zasugerowałoby tożsamości. To, że obiekt niecentralny w naszym polu widzenia może obracać się bez zniekształcenia percepcyjnego, ponieważ jego obraz jest mocno zniekształcony w fizycznej mapie retinotopowej, przedstawia nietrywialny problem obliczeniowy. Ewolucja oszczędza moc obliczeniową poprzez komplikowanie algorytmu. Ewolucja, postrzegana jako presja projektowa, wywiera stałą ekwipotencjalną presję projektową na całą istniejącą złożoność; ludzki programista posługuje się ogólną inteligencją jak skalpelem. Ewolucji nie jest o wiele trudniej „zaprojektować" i „debugować" logarytmiczną mapę wizualną z powodu tej stałej „presji projektowej"; dalsze adaptacje mogą być budowane na podstawie logarytmicznej mapy wizualnej niemal tak samo łatwo, jak na mapie o stałej rozdzielczości. Ogólna inteligencja ludzkiego programisty miałaby trudności ze śledzeniem wszystkich jednoczesnych komplikacji projektowych tworzonych przez mapę logarytmiczną. Może to być możliwe, ale byłoby trudne, zwłaszcza w kontekście badań eksploracyjnych; mapa logarytmiczna przekształca proste problemy projektowe w złożone problemy projektowe, a tym samym przekształca złożone problemy projektowe w koszmary. Zasugerowałbym użycie modalności sensorycznych o stałej rozdzielczości na wczesnych etapach rozwoju sztucznej inteligencji – jak sugerowano powyżej, sugerując modalność sensoryczną modelowaną wokół planszy Go 100x100 – ale implikacją jest to, że te wczesne modalności będą miały niższą rozdzielczość, będą miały mniejsze pole i będą mniej wydajne obliczeniowo. Przeciwny pogląd teoretyczny byłby taki, że złożone ale wydajne modalności wprowadzają niezbędne problemy dla inteligencji. Przeciwstawnym, pragmatycznym poglądem byłoby stwierdzenie, że złożone, ale wydajne modalności łatwiej uwzględnić w dojrzałej sztucznej inteligencji, jeśli zostały uwzględnione w architekturze od samego początku, tak aby uniknąć metaforycznych problemów „Y2K" (wszechobecne zależności od upraszczającego założenia, które później zostaje unieważnione).

Poziom koncepcji

DGI używa terminu koncepcja, aby odnieść się do mentalnych treści leżących u podstaw słów, które łączymy w zdania; koncepcje są kombinatorycznymi blokami konstrukcyjnymi myśli i obrazów mentalnych. Te bloki konstrukcyjne to wyuczona złożoność, a nie wrodzona złożoność; są one abstrahowane z doświadczenia. Struktura koncepcji jest wchłaniana z powtarzających się regularności w postrzeganej rzeczywistości. Koncepcja jest abstrahowana z doświadczeń, które istnieją jako wzorce sensoryczne w jednej lub większej liczbie modalności. Po abstrakcji koncepcja może być porównana do nowego doświadczenia sensorycznego, aby określić, czy nowe doświadczenie spełnia koncepcję, lub równoważnie, czy koncepcja opisuje aspekt doświadczenia. Koncepcje mogą opisywać zarówno środowiskowe doświadczenie sensoryczne, jak i wewnętrznie generowane obrazy mentalne. Koncepcje mogą być również narzucane na bieżące obrazy robocze. W najprostszym przypadku przykład związany z koncepcją może być załadowany do obrazów roboczych, ale konstruowanie złożonych obrazów mentalnych wymaga, aby koncepcja była ukierunkowana na część istniejących obrazów mentalnych, które następnie koncepcja przekształca. Koncepcje są fasetowe; mają one wewnętrzną strukturę i strukturę asocjacyjną, która wchodzi do gry, gdy narzucanie lub opis napotyka przeszkodę na drodze. Faceting może być również wywoływany celowo; na przykład „smakuje jak czekolada" kontra „wygląda jak czekolada". Rozwiązanie któregokolwiek z tych problemów samodzielnie, przy wystarczającym stopniu ogólności i w sposób obliczeniowy, byłoby poważnym

wyzwaniem; rozwiązanie wszystkich trzech problemów jednocześnie stanowi podstawowe wyzwanie zbudowania systemu, który uczy się złożoności w kombinatorycznych fragmentach. „Jądro koncepcji" to pseudosensoryczny wzorzec wytworzony przez abstrahowanie od doświadczenia sensorycznego. Podczas satysfakcji koncepcji jądro to oddziałuje z warstwowymi detektorami cech, aby określić, czy zgłoszony obraz pasuje do ernelu; podczas narzucania koncepcji jądro oddziałuje z warstwowymi kontrolerami cech, aby wytworzyć nowy obraz lub zmienić istniejący obraz. Programista poszukujący dobrej reprezentacji dla jąder koncepcji musi znaleźć reprezentację, która jednocześnie spełnia te wymagania:

1. Reprezentacja jądra może być spełniona i narzucona referentom w modalności sensorycznej.

2. Reprezentacja jądra lub reprezentacja koncepcji zawiera wewnętrzną strukturę potrzebną do fasetowej kombinacji koncepcji, jak w przypadku „trójkątnej żarówki" podanej wcześniej jako przykład.

3. Abstrahowanie nowych reprezentacji jądra przy użyciu doświadczenia sensorycznego jako surowca jest obliczeniowo wykonalne.

Pojęcia mają inne właściwości oprócz złożonych jąder. Jądra wiążą pojęcia z obrazami sensorycznymi, a zatem z poziomem modalności. Pojęcia mają również złożoność, która odnosi się do poziomu pojęć; tj. pojęcia mają złożoność, która wynika z ich relacji do innych pojęć. W Good Old-Fashioned AI ten aspekt pojęć został podkreślony kosztem wszystkich innych, ale nie jest to usprawiedliwienie dla ignorowania relacji pojęcie-pojęcie w nowej teorii. Pojęcia są superkategoriami i podkategoriami siebie nawzajem; istnieją pojęcia, które opisują pojęcia i pojęcia, które opisują relacje między pojęciami. W logice formalnej tradycyjna idea pojęć polega na tym, że pojęcia są kategoriami zdefiniowanymi przez zestaw indywidualnie koniecznych i łącznie wystarczających rekwizytów; że ekstensjonalnym odniesieniem kategorii jest zestaw zdarzeń lub obiektów, które są członkami kategorii; i że połączenie dwóch kategorii jest sumą ich rekwizytów, a zatem przecięciem ich zestawów rekwizytów. Ta formuła jest nieodpowiednia dla złożonej, chaotycznej, nakładającej się struktury kategorii rzeczywistości i jest niezgodna z szerokim zakresem ustalonych efektów poznawczych [57]. Właściwości takie jak zwykle konieczne i zwykle wystarczające wymagania oraz kombinacje pojęć, które są czasami sumą ich wymagań lub przecięciem ich klas ekstensjonalnych, wyłaniają się z podstawowej reprezentacji pojęć – wraz z innymi ważnymi właściwościami, takimi jak efekty prototypowe, w których różnym członkom kategorii przypisuje się różne stopnie typowości. Pojęcia odnoszą się do poziomu myśli przede wszystkim w tym, że są elementami konstrukcyjnymi myśli, ale istnieją również inne skrzyżowania poziomów. Pojęcia introspekcyjne mogą opisywać przekonania i myśli, a nawet rozważania; pojęcie „myśl" jest przykładem. Uogólnienia indukcyjne często dotyczą „pojęć" w tym sensie, że odnoszą się do odniesień pojęcia; na przykład „Trójkątne żarówki są czerwone". Rozważanie może koncentrować się na pojęciu w celu dojścia do wniosków na temat kategorii ekstensjonalnej, a rozważanie introspekcyjne może koncentrować się na pojęciu w jego roli jako obiektu poznawczego. Struktura pojęć jest wszechobecna w procesach percepcyjnych i poznawczych, ponieważ struktura kategorii jest wszechobecna w procesach niskiej entropii naszego wszechświata o niskiej entropii.

Istota pojęć

Jednym ze znaczeń słowa „abstrakcja" jest „usuwanie"; w chemii abstrakcja atomu oznacza odjęcie go od grupy cząsteczkowej. Użycie terminu „abstrakcja" do opisania procesu tworzenia pojęć można rozumieć jako implikację dwóch poglądów: Po pierwsze, że stworzenie pojęcia oznacza uogólnienie; po drugie, że uogólnienie oznacza utratę informacji. Abstrakcja jako utrata informacji jest klasycznym poglądem na pojęcia (tj. poglądem na pojęcia w ramach GOFAI i logiki formalnej). Tworzenie pojęcia „czerwony" jest rozumiane jako skupianie się tylko na kolorze, kosztem innych cech, takich jak rozmiar i kształt; uważa się, że wszelkie wykorzystanie pojęć polega na celowej utracie informacji. Problem z

klasycznym poglądem polega na tym, że dopuszcza on jedynie ograniczony repertuar pojęć. To prawda, że niektóre pojęcia najwyraźniej prowadzą do prostej utraty informacji. Zadanie dotarcia do jądra pojęcia dla pojęcia „czerwony" – jądra zdolnego do interakcji z obrazami wizualnymi w celu odróżnienia obiektów czerwonych od obiektów nieczerwonych – jest stosunkowo trywialne. Nawet jednoczesne spełnienie problemów abstrakcji i satysfakcji dla „czerwonego" jest stosunkowo trywialne. Wystarczą dobrze znane, w pełni ogólne narzędzia, takie jak sieci neuronowe lub obliczenia ewolucyjne. Aby nauczyć się rozwiązywać problem satysfakcji, sieć neuronowa musi jedynie nauczyć się wyzwalać, gdy detektory cech na poziomie modalności dla „koloru" zgłaszają określony kolor – punkt mieszczący się w określonej objętości przestrzeni kolorów – na szerokim obszarze, a w przeciwnym razie nie wyzwalać. Fragment kodu musi ewoluować jedynie w celu przetestowania tej samej cechy. (Sieć neuronowa prawdopodobnie trenowałaby szybciej do tego zadania). Wystarczająco wyrafinowana modalność uprościłaby to zadanie jeszcze bardziej, wykonując większość pracy grupowania obrazów wizualnych w obiekty i wykrywając powierzchnie o jednolitym kolorze, tym samym odcieniu lub w większości tego samego odcienia. Ludzka modalność wzrokowa idzie jeszcze dalej i wstępnie kategoryzuje kolory, dzieląc je na złożoną przestrzeń barw [7], przestrzeń barw ma jedenaście kulturowo uniwersalnych objętości ogniskowych [4], przy czym objętości ogniskowe mają stosunkowo ostre granice wewnętrzne w stosunku do fizycznie ciągłych zmian długości fali (patrz [93] lub po prostu spójrz na pasma w tęczy). Rozróżnianie w obrębie wrodzonych granic kolorów jest łatwe; rozróżnianie w granicach kolorów jest trudne [68]. Tak więc ludzka modalność wzrokowa dostarcza bardzo silnych sugestii co do tego, gdzie leżą granice w przestrzeni barw, chociaż nadal wymagany jest ostatni krok kategoryzacji. Biorąc pod uwagę modalność wzrokową, koncepcja czerwieni leży bardzo blisko metaforycznej „powierzchni" modalności. U ludzi czerwień prawdopodobnie znajduje się na powierzchni, jest bezpośrednim wyjściem detektorów cech modalności. W AI z mniej wyrafinowanymi modalnościami wzrokowymi „czerwień" jako kategoria musiałaby zostać abstrakcyjnie przedstawiona jako rozmyta objętość w gładkiej przestrzeni barw pozbawionej ludzkich granic. Czerwone jądro koncepcji (u ludzi i sztucznej inteligencji) musi być bardziej złożone niż prosty test binarny lub test klastrowania rozmytych kolorów, ponieważ „czerwień", jak rozumiemy, opisuje obszary wizualne, a nie pojedyncze piksele (chociaż czerwień może opisywać „obszar wizualny" składający się z małego punktu). Mimo to złożoność związana z koncepcją czerwieni leży niemal całkowicie w modalności sensorycznej, a nie w jądrze koncepcji. Takie koncepcje moglibyśmy nazwać koncepcjami powierzchniowymi. Nawet w przypadku koncepcji powierzchniowych jednoczesne rozwiązywanie abstrakcji, satysfakcji i narzucania byłoby prawdopodobnie o wiele bardziej wykonalne przy użyciu specjalnej reprezentacji jąder koncepcji, a nie generycznie wyszkolonych sieci neuronowych lub programów ewolucyjnych. Nakładanie wymaga jądra koncepcji, które można selektywnie stosować do obrazowania w ramach modalności wizualnej, przekształcając to obrazowanie tak, aby końcowy wynik spełniał koncepcję. W przypadku koncepcji „czerwonej" jądro koncepcji oddziaływałoby z kontrolerami cech dla koloru, a docelowe obrazy mentalne stałyby się czerwone. Nie można tego zrobić, malując każdy pojedynczy piksel tym samym odcieniem czerwieni; taka transformacja zatarłaby krawędzie, powierzchnie, tekstury i wiele innych cech wysokiego poziomu, które intuicyjnie powinny zostać zachowane. Wizualizacja „czerwonej cytryny" nie powoduje, że umysł wyobraża sobie jasnoczerwoną plamę z konturem cytryny. Jądro koncepcji nie wysyła oddzielnych poleceń koloru do kontrolera cech niskiego poziomu każdego indywidualnego elementu wizualnego; raczej jądro koncepcji narzuca czerwień w połączeniu z innymi aktualnie aktywowanymi cechami, aby przedstawić czerwoną cytrynę, która zachowuje krawędź, kształt, krzywiznę powierzchni, teksturę i inne wizualizowane cechy początkowego obrazu cytryny. Prawdopodobnie dzieje się tak, ponieważ postrzegane zabarwienie jest właściwością powierzchni i obiektów wizualnych, a nie, lub jak również, poszczególnych elementów wizualnych, a nasze jądro koncepcji czerwieni oddziałuje z tą cechą wysokiego poziomu, która następnie rozchodzi się falowo w spójnym połączeniu z innymi cechami. Abstrahowanie narzucalnego

jądra koncepcji dla „czerwonego” jest problemem o innym zakresie niż abstrahowanie zadowalającego jądra dla „czerwonego”. Istnieje natychmiast oczywisty sposób trenowania sieci neuronowej w celu wykrywania satysfakcji z „czerwonego”, biorąc pod uwagę zestaw treningowy znanych doświadczeń „czerwonego” i nie-„czerwonego”, ale nie ma równie oczywistej procedury nauczania dla problemu narzucania „czerwonego”. Najbardziej bezpośrednią miarą sukcesu jest stopień, w jakim przekształcone obrazy spełniają wymagania sieci neuronowej już wyszkolonej w zakresie wykrywania „czerwonego”, ale jasnoczerwona plama w kształcie cytryny prawdopodobnie będzie bardziej „czerwona” niż wizualizowana czerwona cytryna. W jaki sposób jądro dochodzi do transformacji, która wprowadza spójną zmianę w zabarwieniu obiektu, a nie transformacji, która maluje wszystkie elementy wizualne na nieokreślony odcień czerwieni lub transformacji, która ładuje losowy czerwony obiekt do pamięci? Każda z tych transformacji spełniałaby koncepcję „czerwonego”. Można by wytrenować w pełni ogólne sieci neuronowe, aby narzucały minimalne transformacje, choć nie jestem pewien, czy „minimalna transformacja” jest regułą, która powinna rządzić narzucaniem koncepcji. Niezależnie od rzeczywistej podatności tego problemu na rozwiązywanie, mocno wątpię, aby ludzkie systemy poznawcze tworzyły koncepcje poprzez trenowanie ogólnych sieci neuronowych w zakresie satysfakcji i narzucania. Podejrzewam, że koncepcje nie mają niezależnych procedur dla satysfakcji i narzucania; podejrzewam również, że ani satysfakcja, ani narzucanie nie są produktem uczenia się wzmacniającego w pełni ogólnej procedury. Podejrzewam raczej, że jądro koncepcji składa się ze wzorca w reprezentacji powiązanego z (ale nie identycznego z) reprezentacją obrazów sensorycznych, że ten wzorzec jest wytwarzany przez transformację doświadczeń, z których koncepcja jest abstrahowana, i że ten wzorzec oddziałuje z modalnością w celu wdrożenia zarówno satysfakcji koncepcji, jak i narzucania koncepcji. Bardzo prostym przykładem nieproceduralnego, opartego na wzorcach jądra koncepcji byłoby „grupowanie na pojedynczej cesze”. Czerwień można by wyabstrahować z bazy doświadczalnej poprzez obserwację niezwykłego grupowania wartości punktowych dla cechy koloru. Załóżmy, że AI zostaje wyzwana wirtualną grą, w której celem jest znalezienie „kluczy” do „zamka” poprzez wybranie obiektów z dużego zestawu próbek. Kiedy AI pomyślnie przejdzie pięć prób, wybierając prawidłowy obiekt za pierwszym razem, przyjmuje się, że AI nauczyła się reguły. Załóżmy, że regułą gry jest to, że „czerwone” obiekty otwierają zamek, a AI zgromadziła już bazę doświadczeń z poprzednich porażek i sukcesów w poszczególnych próbach. Zakładając użycie trójwymiarowej przestrzeni kolorów, wartości kolorów prawidłowych kluczy reprezentowałyby ścisły klaster w stosunku do rozkładu wśród wszystkich potencjalnych kluczy. Stąd abstrakcyjne jądro koncepcji mogłoby przybrać formę pary cecha-klaster, gdzie cechą jest kolor, a klaster jest punktem centralnym plus pewna miara odchylenia standardowego. Tworzy to jądro koncepcji z prototypem i ilościową spełnialnością; koncepcja ma punkt centralny i rozmyte, ale rzeczywiste granice. To samo jądro koncepcji może być również narzucone na wybrany fragment obrazu mentalnego poprzez załadowanie centralnego punktu koloru do kontrolera cech koloru – to znaczy załadowanie wartości klastra do kontrolera cech odpowiadającego detektorowi cech klastrowanych. Klastrowanie tego typu ma również pośrednie implikacje dla relacji koncepcja-koncept: „objętość koloru” czerwonego pojęcia może nakładać się na pobliskie pojęcie, takie jak burgund, lub może okazać się obejmować to pojęcie; fakt na poziomie modalności, który z czasem może naturalnie prowadzić do relacji asocjacyjnej lub relacji superkategorii na poziomie koncepcji. Nie nastąpiłoby to u ludzi poprzez bezpośrednie porównanie reprezentacji jąder koncepcji, ale poprzez obserwację nakładania się lub włączenia w kategorie odniesień ekstensjonalnych. Bardziej introspektywna sztuczna inteligencja mogłaby czasami skorzystać z inspekcji reprezentacji jądra, ale powinno to być dodatkiem do doświadczalnego wykrywania relacji kategorii, a nie jego substytutem. Klastrowanie na pojedynczej cesze zdecydowanie nie jest kompletnym systemem koncepcyjnym. Klastrowanie pojedynczej cechy nie jest w stanie zauważyć korelacji między dwiema cechami, jeśli żadna z nich nie jest klastrowana sama; klasteryzacja pojedynczej cechy nie może w żaden sposób korelować krzyżowo

dwóch cech. Koncepcje, które są ograniczone do klasteryzacji pojedynczej cechy, zawsze będą ograniczone do koncepcji na bezpośredniej powierzchni danej modalności sensorycznej. Jednocześnie system pojęć nie jest ogólną inteligencją i nie musi być zdolny do reprezentowania każdej możliwej relacji. Załóżmy, że człowiek został wyzwany do gry, w której „poprawny klucz" zawsze miał kolor, który leżał na dokładnej powierzchni kuli w przestrzeni kolorów; czy ludzki system tworzenia pojęć mógłby bezpośrednio abstrahować tę właściwość? Przypuszczam, że nie; przypuszczam, że co najwyżej człowiek mógłby zauważyć, że klucz ma tendencję do przynależności do pewnej grupy kolorów; tj. mógłby podzielić powierzchnię tej kuli kolorów na oddzielne obszary i założyć, że klucze rozwiązań należą do jednego z kilku obszarów kolorów. Tak więc, nawet jeśli w tym przypadku podstawowa „reguła" jest obliczeniowo bardzo prosta, mało prawdopodobne jest, aby człowiek stworzył koncepcję, która bezpośrednio uwzględnia regułę; może być nawet niemożliwe, aby człowiek mógł wyabstrahować jądro, które wykonuje to proste obliczenie. System tworzenia pojęć nie musi być ogólnie inteligentny sam w sobie; nie musi reprezentować wszystkich możliwych prawidłowości percepcyjnych; wystarczy, aby umysł mógł działać. Podejrzewam, że projekt systemu używany przez ludzi i dobry projekt dla AI okażą się repertuarem różnych metod formowania pojęć. („Klastrowanie na pojedynczej cesze" może być jedną z takich metod lub może być szczególnym przypadkiem bardziej ogólnej metody). Faceting pojęć może być wtedy wynikiem pojęć z wieloma jądrami, tak że pojęcie wykorzystuje więcej niż jedną metodę kategoryzacji w stosunku do swoich odniesień percepcyjnych, lub wewnętrznej struktury w jednym jądrze, lub obu. Jeśli niektóre aspekty odniesień percepcyjnych są bardziej wyraziste, wówczas jądra, które pasują do tych aspektów, prawdopodobnie będą miały większą wagę w pojęciu. Faceting w pojęciu, wynikający z wielu nierównych jąder lub faceting w jednym złożonym jądrze, wydaje się najbardziej prawdopodobnym źródłem efektów prototypu w kategorii.

Etapy procesów koncepcyjnych

Tworzenie koncepcji to proces wieloetapowy. Aby sztuczna inteligencja mogła utworzyć nową koncepcję, musi ona mieć odpowiednie doświadczenia, percepcyjnie grupować doświadczenia, zauważać możliwe ukryte podobieństwa wśród członków grupy (może to być to samo postrzegane podobieństwo, które doprowadziło do pierwotnego grupowania doświadczalnego), weryfikować generalizację, inicjować nową koncepcję jako wyróżnioną treść poznawczą, tworzyć jądro(a) koncepcji poprzez abstrakcję z bazy doświadczalnej i integrować nową koncepcję z systemem. (Ta lista kontrolna jest przeznaczona jako tymczasowe przybliżenie; rzeczywiste projekty umysłu mogą się różnić, ale prawdopodobnie nadal będzie zaangażowana sekwencja czasowa). W podanym wcześniej przykładzie sztuczna inteligencja abstrahuje czerwień, zaczynając od zdarzenia oddolnego, napędzanego doświadczeniem: zauważając możliwe skupienie cechy koloru w obrębie wcześniej istniejących kluczy kategorii. Można by przypuszczać, że sprawdzanie klastrowania kolorów mogło zostać zasugerowane odgórnie, na przykład przez jakieś heurystyczne przekonanie, ale w tym przykładzie założymy, że pierwotne postrzeganie podobnego ubarwienia było nieoczekiwanym, oddolnym zdarzeniem; produktem ciągłych i automatycznych sprawdzeń klastrowania pojedynczej cechy wśród wszystkich cech wysokiego poziomu w obecnie istotnych kategoriach doświadczalnych. Zamiast być częścią istniejącego ciągu myśli, wykrycie klastrowania tworzy zdarzenie „Aha!", nowe zdarzenie poznawcze o wysokiej wyrazistości, które staje się przedmiotem uwagi, tymczasowo odsuwając na bok poprzedni ciąg myśli. (Zobacz dyskusję na temat poziomu myśli.)

Jeśli skanowanie klastrowania i innych kategoryzowanych podobieństw jest ciągłym zadaniem w tle, może to oznaczać znaczne wydatkowanie zasobów obliczeniowych – być może znaczny procent mocy obliczeniowej wykorzystywanej przez AI. Jest to prawdopodobnie cena posiadania procesu poznawczego, który może być napędzany zarówno przez przerwania oddolne, jak i sekwencje odgórne,

oraz cena posiadania procesu poznawczego, który może czasami zauważać to, co nieoczekiwane. Stąd wydajność, optymalizacja i skalowalność algorytmów dla takich ciągłych zadań w tle mogą odgrywać główną rolę w określaniu wydajności AI. Jeśli obrazowanie pozostanie na miejscu wystarczająco długo, spekulowałbym, że możliwe będzie przekazanie zadania zauważenia możliwego klastrowania do odległych części rozproszonej sieci, podczas gdy zadanie weryfikacji klastrowania i wszystkich kolejnych działań poznawczych pozostanie w lokalnym procesie. Większość mocy obliczeniowej jest wymagana do znalezienia wskazówki, a nie do weryfikacji dopasowania, a fałszywa wskazówka nie wyrządza żadnych szkód (zakładając, że fałszywe wskazówki nie są złośliwymi atakami z niezaufanych węzłów). Gdy podejrzenie podobieństwa zostanie wywołane przez wskazówkę wychwyconą przez ciągły proces w tle, a rzeczywisty stopień podobieństwa zostanie zweryfikowany, AI będzie w stanie stworzyć koncepcję jako treść poznawczą. W powyższym przykładzie proces, który zauważa możliwe klastrowanie, jest zasadniczo tym samym procesem, który weryfikuje klastrowanie i oblicza stopień klastrowania, środek klastrowania i wariancję w klastrze. Tak więc, grupowanie na pojedynczej cesze może skompresować do jednego etapu wskazywanie, opis i abstrakcję podstawowego podobieństwa. Biorąc jednak pod uwagę koszt ciągłego procesu tła, podejrzewam, że zazwyczaj najlepiej będzie oddzielić tańszy mechanizm wskazywania jako proces tła i użyć tego mechanizmu wskazywania do sugerowania bardziej szczegółowych i kosztownych skanów. (Należy zauważyć, że jest to „równoległe skanowanie tarasowe"; patrz [84] i [38].) Po utworzeniu koncepcji i jądra(ów) koncepcji, sztuczna inteligencja mogłaby zauważyć relacje koncepcja-pojęcie, takie jak relacje superkategorii i podkategorii. Nie wierzę, że relacje koncepcja-pojęcie są obliczane przez bezpośrednie porównywanie reprezentacji jądra; myślę, że relacje koncepcja-pojęcie są poznawane przez uogólnianie na całe wykorzystanie koncepcji. Może to być dobra heurystyka, aby szukać relacji koncepcja-pojęcie natychmiast po utworzeniu nowej koncepcji, ale byłby to oddzielny utwór w ramach rozważań, a nie automatyczna część tworzenia koncepcji. Po uformowaniu koncepcji, nowa koncepcja musi zostać zintegrowana z systemem. Abyśmy mogli uznać, że koncepcja została naprawdę „zintegrowana z systemem" i teraz przyczynia się do inteligencji, koncepcja musi zostać wykorzystana. Skanowanie całej zapisanej bazy koncepcji w celu znalezienia koncepcji, które są spełnione przez obecne wyobrażenia mentalne, obiecuje być procesem jeszcze bardziej kosztownym obliczeniowo niż ciągłe sprawdzanie przeszłości pod kątem klastrowania. Indywidualna kontrola satysfakcji jest prawdopodobnie mniej intensywna obliczeniowo niż przeprowadzanie narzucania koncepcji – ale kontrole satysfakcji wydają się prawdopodobnie ciągłą operacją w tle, przynajmniej u ludzi. Jak omówiono wcześniej, ludzie i SI mają różne podłoża obliczeniowe: ludzie są powolni, ale ogromnie równoległi; SI są szybkie, ale ubogie w zasoby. Jeśli ludzie okażą się rutynowo paralelizowali względem wszystkich nauczonych koncepcji, SI może po prostu nie być w stanie sobie na to pozwolić. Optymalne rozwiązanie SI może obejmować porównywanie obrazów roboczych z mniejszym podzbiorem nauczonej złożoności – tylko kilka koncepcji, przekonań lub wspomnień byłoby skanowanych względem obrazów roboczych w dowolnym momencie. Alternatywnie, SI może być w stanie użyć skanowania tarasowego, rozmytego hashowania lub sortowania rozgałęzionego, aby uczynić problem wykonalnym. Jednym z obiecujących znaków jest zjawisko poznawczego przygotowania do powiązanych koncepcji, które sugeruje, że ludzie, pomimo swojego paralelizmu, nie używają czystej siły brutalnej. Niezależnie od tego, przypuszczam, że dopasowywanie obrazów do dużych zestawów koncepcji będzie jednym z najbardziej intensywnych obliczeniowo podprocesów w AI, być może najdroższym podprocesem. Dopasowywanie koncepcji jest zatem kolejnym dobrym kandydatem do dystrybucji w ramach „zauważania na odległość, weryfikacji lokalnej"; należy również zauważyć, że baza koncepcji może zostać podzielona na rozproszone procesory, chociaż może to uniemożliwić algorytmom dopasowującym wykorzystywanie regularności w bazie koncepcji i procesie dopasowywania.

Złożone koncepcje i struktura „Pięciu"

W klasycznej filozofii abstrakcji kategorii, abstrakcja polega wyłącznie na selektywnym skupieniu się na informacjach, które są już znane; skupieniu się na „kolorze" lub „czerwieni" obiektu, w przeciwieństwie do jego kształtu, położenia lub prędkości. W „jądrach pojęć" DGI wewnętrzna reprezentacja pojęcia ma złożoność wykraczającą poza utratę informacji – nawet w przypadku „czerwieni" i innych pojęć, które leżą niemal bezpośrednio na powierzchni modalności sensorycznej. Jedynym pojęciem, które jest czystą utratą informacji, jest pojęcie, które leży całkowicie na powierzchni modalności; pojęcie, którego zadowolenie jest dokładnie równe zadowoleniu pewnego pojedynczego detektora cech. Pojęcie „czerwieni", opisane wcześniej, jest w rzeczywistości rozmytym perceptem stopni czerwieni. Biorąc pod uwagę, że AI ma płaską przestrzeń kolorów, a nie ludzką przestrzeń kolorów z wrodzonymi ogniskowymi objętościami i granicami kolorów, percept „czerwieni" zawierałby co najmniej tyle dodatkowej złożoności – ponad złożonością na poziomie modalności – ile jest używane do opisu klastrowania. Na przykład „grupowanie na pojedynczej cesze" może przybrać formę opisu rozkładu Gaussa wokół punktu centralnego. Konkretne użycie rozkładu Gaussa nie przyczynia się do użytecznej inteligencji, chyba że środowisko również wykazuje grupowanie Gaussa, ale rozkład Gaussa jest prawdopodobnie przydatny, aby umożliwić AI zauważenie szerokiej klasy grupowań wokół punktu centralnego, nawet grupowań, które w rzeczywistości nie podążają za rozkładem Gaussa. Nawet w przypadku braku bezpośredniej regularności środowiskowej, koncepcja może przyczynić się do skutecznej inteligencji, umożliwiając postrzeganie bardziej złożonych regularności. Na przykład naprzemienna sekwencja kluczowych obiektów „czerwonych" i „zielonych" może nie przejść testów klasteryzacji na poziomie modalności, ponieważ żaden klaster Gaussa nie zawiera (prawie) wszystkich sukcesów i wyklucza (prawie) wszystkie porażki. Jednak jeśli AI wcześniej opracowała koncepcje dla „czerwonego" i „zielonego", naprzemienne powtarzanie spełnienia koncepcji „czerwonych" i „zielonych" jest potencjalnie wykrywalne przez detektory powtórzeń wyższego poziomu. Podział przestrzeni barw na koncepcje powierzchniowe sprawia, że wykrywanie przemienności wyższego rzędu staje się obliczeniowo wykonalne. Nawet tworzenie prostych koncepcji – koncepcji leżących na powierzchni modalności – rozszerza możliwości percepcyjne AI i zakres problemów, które AI może rozwiązać. Koncepcje mogą również ucieleśniać regularności, które nie są bezpośrednio reprezentowane w żadnej modalności sensorycznej i które nie są żadną kowariancją ani klastrowaniem detektorów cech już znajdujących się w modalności sensorycznej. Program „Copycat" Melanie Mitchell i Douglasa Hofstadtera działa w domenie ciągów liter, takich jak „abc", „xyz", „onml", „ddd", „cwj" itd. Funkcją Copycat jest dokończenie problemów analogii, takich jak „abc:abd::ace:?" [37]. Ponieważ Copycat jest modelem tworzenia analogii percepcyjnych, a nie modelem tworzenia kategorii, Copycat ma ograniczony zasób wstępnie zaprogramowanych koncepcji i nie uczy się dalszych koncepcji poprzez doświadczenie. (Nie należy tego traktować jako krytyki projektu Copycat; badacze wyraźnie zauważyli, że nie badano formowania pojęć). Załóżmy, że ogólna sztuczna inteligencja (nie Copycat), pracująca w dziedzinie zabawek ciągów liter, napotyka problem, który można rozwiązać tylko poprzez odkrycie, co sprawia, że ciągi liter „hcfrb", „yhumd", „exbvb" i „gxqrc" są podobne do siebie, ale niepodobne do ciągów „ndaxfw", „qiqa", „r", „rvm" i „zinw". Copycat ma wbudowaną zdolność do zliczania liter w ciągu lub grupie; w terminologii DGI Copycat można by powiedzieć, że wyodrębnia liczbę jako cechę na poziomie modalności. Istnieją liczne dowody na to, że ludzie również mają wsparcie oprogramowania mózgowego dla subityzacji (bezpośredniego postrzegania) małych liczb i wsparcie oprogramowania mózgowego dla postrzegania przybliżonych ilości dużych liczb (przegląd można znaleźć w [20]). Załóżmy jednak, że ogólna SI nie posiada zdolności liczenia na poziomie modalności. W jaki sposób SI mogłaby utworzyć kategorię „pięć" lub nawet „grupy pięciu liter"? To wyzwanie wskazuje na inherentny deficyt punktu widzenia „utraty informacji" w abstrakcji. W przypadku SI bez wsparcia subityzacyjnego – lub w przypadku człowieka, który ma problem z liczbą taką jak „dziewięć", która jest poza zasięgiem ludzkiej subityzacyjnej – wyróżniająca cecha, kardynalność, nie jest reprezentowana przez modalność (lub u ludzi reprezentowana jest tylko w przybliżeniu). Zarówno w

przypadku ludzi, jak i SI, zdolność do tworzenia pojęć dla niesubityzowalnych dokładnych liczb wymaga czegoś więcej niż zdolności do selektywnego skupiania się na aspekcie „liczby", a nie na aspekcie „lokalizacji" lub „litery" (lub „koloru", „kształtu" lub „wysokości"). Podstawowym wyzwaniem nie jest skupianie się na aspekcie liczbowym, ale raczej postrzeganie „aspektu liczbowego" w pierwszej kolejności. Na potrzeby tej dyskusji nie mówimy o zdolności rozumienia liczb, arytmetyki lub matematyki, tylko o zdolności SI do tworzenia kategorii „pięć". Posiadanie kategorii „pięć" nie oznacza nawet posiadania kategorii „cztery" lub „sześć", a tym bardziej sformułowania abstrakcyjnej superkategorii „liczba". Podobnie „odkrycie" piątki nie jest uważane za matematycznie znaczące. W terminologii matematycznej niemal każdy zestaw poznawczych bloków konstrukcyjnych wystarczy do odkrycia liczb; liczby są fundamentalne i można je skonstruować za pomocą szerokiej gamy różnych procedur powierzchniowych. Znaczącym osiągnięciem nie jest „wyciskanie" liczb z systemu tak rozproszonego, że najwyraźniej brakuje mu zwykłych prekursorów liczby. Raczej wyzwaniem jest przedstawienie opisu odkrycia „pięćki" w sposób, który uogólnia również odkrycie innych złożonych pojęć. Hipotetyczne bloki konstrukcyjne pojęcia powinny być ogólne (przydatne w budowaniu innych, nienumerycznych pojęć), a hipotetyczne relacje między blokami konstrukcyjnymi powinny być ogólne. Dopuszczalne jest, aby odkrycie „pięćki" było proste, ale metoda odkrycia musi być ogólna. Działająca, ale prymitywna procedura spełniania koncepcji „piątki", po odkryciu piątki, może wyglądać mniej więcej tak: Skup się na grupie docelowej (grupie, która może lub nie spełniać „piątki"). Przywołaj z pamięci wzór dla „piątki" (czyli jakieś konkretne przeszłe doświadczenie, które stało się wzorem dla koncepcji „piątki"). Wyobraź sobie wzór „piątki" w oddzielnej przestrzeni roboczej. Narysuj korespondencję z obiektu w grupie, która jest wzorem pięciu, do obiektu w grupie, która jest celem. Powtarzaj tę procedurę, aż nie pozostanie żaden obiekt w obrazie przykładu lub nie pozostanie żaden obiekt w obrazie docelowym. Nie rysuj korespondencji z jednego obiektu do drugiego, jeśli taka korespondencja już istnieje. Jeśli po zakończeniu tej procedury nie będzie żadnych zwisających obiektów w przykładzie lub w grupie docelowej, oznacz grupę docelową jako spełniającą koncepcję „piątki". W tym przykładzie własność „pięć" tłumaczy się na własność: „Mogę skonstruować kompletne odwzorowanie, bez żadnych zwisających elementów, używając unikalnych odpowiedniości, między tą docelową grupą obiektów a pewną grupą obiektów, których obraz mentalny odzyskałem z pamięci". Jest to matematycznie proste, ale poznawczo ogólne. Na poparcie tezy, że „odpowiedniość", „unikalna odpowiedniość" i 436 Eliezer Yudkowsky „kompletne odwzorowanie bez żadnych zwisających elementów" są ogólnymi prymitywami pojęciowymi, a nie konstrukcjami przydatnymi wyłącznie do odkrywania liczb, należy zauważyć, że Copycat obejmuje odpowiedniości, unikalne odpowiedniości i percepcyjny pęd w kierunku kompletnych odwzorowań [71]. Copycat ma bezpośrednią proceduralną implementację zmysłu liczby i nie używa tych konstrukcji odwzorowania do budowania pojęć liczbowych. Konstrukcje odwzorowania, które przywołałem dla liczby, są niezależnie niezbędne dla teorii Copycata dotyczącej tworzenia analogii jako percepcji. Gdy procedura kończy się etykietowaniem obrazów koncepcją „pięć", obrazy te stają się doświadczalnym przykładem koncepcji „pięć". Jeśli przykłady powiązane z proceduralnie zdefiniowaną koncepcją mają jakieś uniwersalne cechy lub częste cechy, które są zauważalne percepcyjnie, koncepcja może nabyć jądra po fakcie, chociaż jądro może wyrażać się jako wskazówka lub oczekiwanie, a nie być koniecznym i wystarczającym warunkiem spełnienia koncepcji. Koncepcje z proceduralnymi definicjami są regularnymi koncepcjami i mogą posiadać jądra, przykłady, skojarzone wspomnienia itd. Jaka jest korzyść z rozkładania „pięć" na złożoną procedurę, zamiast po prostu pisać kodlet lub detektor cech na poziomie modalności, który bezpośrednio liczy (subiektuje) członków grupy? Podstawowym powodem preferowania rozwiązania niemodalnego w tym przykładzie jest wykazanie, że SI musi być w stanie rozwiązywać problemy, których nie przewidziano podczas projektowania. Z tej perspektywy „pięć" jest złym przykładem do wykorzystania, ponieważ jest bardzo mało prawdopodobne, aby programista SI nie przewidział numeryczności podczas fazy projektowania. Jednak rozkładalna koncepcja dla „piątki"

i detektor cech na poziomie modalności, który podstawia wszystkie liczby do (232-1), mogą być również porównywane pod względem tego, jak dobrze wspierają ogólną inteligencję. Pomimo znacznie większego narzutu obliczeniowego, argumentowałbym, że rozkładalna koncepcja jest lepsza od detektora cech na poziomie modalności. Modalność bilardowa z detektorem cech, który podstawia wszystkie kule bilardowe w percepcyjnej grupie i wyprowadza percepcyjnie odrębną etykietę – „detektor numeronów" – wystarczy do rozwiązania wielu natychmiastowych problemów, które wymagają wyczucia liczby. Jednakże SI, która używa tego detektora cech do utworzenia koncepcji powierzchni dla „pięciu", nie będzie w stanie subityzować „pięciu" grup bilardowych w ramach supergrupy, chyba że programista miał również na tyle dalekowzroczności, aby rozszerzyć detektor cech subityzujących na zliczanie grup, a także określonych obiektów. Podobnie, ta uniwersalna zdolność subityzowania nie będzie obejmować wielu modalności, chyba że programista miał na tyle dalekowzroczności, aby rozszerzyć tam również detektor cech. Brainware jest ograniczony do tego, o czym programista myślał w danym momencie. Czy SI rozumie „pięćdziesiątkę", gdy staje się w stanie policzyć pięć jabłek? Albo gdy SI może również policzyć pięć zdarzeń w dwóch różnych modalnościach? Albo gdy SI może policzyć pięć własnych myśli? Rozszerzenie detektora cech o obsługę dowolnego z tych przypadków jako przypadku szczególnego jest programowo trywialne, ale jest to ścieżka, która kończy się wymaganiem nieskończonej ilości majsterkowania w celu wdrożenia rutynowych procesów myślowych (tj. nierozkładalność powoduje „problem zdrowego rozsądku"). Najważniejszym powodem dekompozycji jest to, że koncepcje o zorganizowanych strukturach wewnętrznych są bardziej zmienne. Detektor numeronów zaprogramowany przez człowieka, zmutowany na poziomie kodu, prawdopodobnie po prostu by się zepsuł. Koncepcja o strukturze wewnętrznej lub proceduralnej, stworzona przez własne procesy myślowe SI w odpowiedzi na doświadczenie, jest zmienna przez procesy myślowe SI w odpowiedzi na dalsze doświadczenie. Na przykład Douglas Lenat potwierdza , że najtrudniejszą częścią tworzenia Eurisko było wynalezienie dekompozycji reprezentacji dla heurystyk, tak aby klasa transformacji dostępna dla Eurisko czasami skutkowała ulepszeniami, a nie zepsutymi fragmentami kodu i błędami LISP. Opisanie tego jako gładkich krajobrazów sprawności prawdopodobnie zbyt mocno rozciąga metaforę, ale „wygładzanie" w jakiejś formie jest zdecydowanie zaangażowane. Surowy kod ma tylko jeden poziom organizacji, a zmiana losowej instrukcji na tym poziomie zwykle po prostu psuje całą funkcję. Heurystyka Eurisko została podzielona na fragmenty i mogła być manipulowana (za pomocą heurystyki Eurisko) na poziomie fragmentu. Lokalne przesunięcia w fragmentach procedury „pięć" dają wiele przydatnych potomków. Selektywnie łagodząc wymóg „żadnych zwisających obiektów" w obrazie docelowym, otrzymujemy koncepcję „mniej niż lub równe pięć"-ności. Łagodząc wymóg „żadnych zwisających obiektów" w obrazie wzorcowym, otrzymujemy koncepcję „większe niż lub równe pięć"-ności. Wymagając jednego lub więcej zwisających obiektów w obrazie docelowym, otrzymujemy koncepcję „więcej niż pięć"-ności. Porównując dwa obrazy docelowe, zamiast przykładu i obrazu, otrzymujemy koncepcję „odpowiedniości jeden do jednego między członkami grupy" (co nazwalibyśmy „tą samą liczbą co" w innej procedurze), a stamtąd „mniej niż" lub „mniej niż lub równe" i tak dalej. Jedno z tych pojęć, jednoznaczna korespondencja między dwoma obrazami mentalnymi, nie jest po prostu użytecznym potomstwem pojęcia „pięćdziesiątki", ale prostszym potomstwem. Tak więc prawdopodobnie wcale nie jest „potomstwem", ale koncepcją wstępną, która sugeruje realną ścieżkę do zrozumienia piątki. Wiele zadań fizycznych w naszym świecie wymaga równych liczb (odpowiadających zestawów) dla pewnej grupy; cztery kołki dla czterech otworów, dwa buty dla dwóch stóp.

Doświadczalne ścieżki do złożonych koncepcji

Rozważmy zadanie w świecie rzeczywistym polegające na umieszczeniu czterech kołków w czterech otworach. Kołek nie może wypełnić dwóch otworów; dwa kołki nie zmieszczą się w jednym otworze. Stałe obiekty nie mogą zajmować tego samego miejsca, nie mogą pojawiać się w wielu miejscach

jednocześnie i nie pojawiają się ani nie znikają spontanicznie. Te reguły środowiska fizycznego znajdują odzwierciedlenie w domyślnych zachowaniach naszej własnej modalności wzrokowo-przestrzennej; nawet wczesne niemowlęta przedstawiają obiekty jako ciągłe i będą dłużej patrzeć na sceny, które implikują naruszenia ciągłości. Na podstawie problemów ze świata rzeczywistego, takich jak kołki i otwory, lub ich odpowiedników w mikroświecie, sztuczna inteligencja może rozwijać koncepcje, takie jak unikalna korespondencja: kołek nie może wypełnić wielu otworów, wiele kołków nie zmieści się w jednym otworze. Sztuczna inteligencja może nauczyć się reguł rysowania unikalnej korespondencji i testować reguły na podstawie doświadczenia, zanim napotka potrzebę utworzenia bardziej złożonej koncepcji „pięćdziesiątki". Obecność natychmiastowego, lokalnego testu użyteczności oznacza, że zaobserwowane niepowodzenia i sukcesy mogą jednoznacznie przyczynić się do utworzenia koncepcji, która jest „prosta" w stosunku do już wyszkolonej bazy koncepcji. Jeśli nowa koncepcja zawiera wiele nowych, nieprzetestowanych części i wystąpi błąd, wówczas dla AI może nie być jasne, który lokalny błąd spowodował globalną awarię. Jeśli AI spróbuje podzielić „pięć" na kawałki w jednym kroku, a obecna procedura spełnienia „pięć" zawiedzie — zostanie pozytywnie spełniona przez grupę inną niż piątka lub niezaspokojona przez grupę piątka — dla AI może nie być jasne, że globalna awaria wynikała z lokalnego błędu niejednoznacznej korespondencji. Pełna ścieżka do piątki prawdopodobnie obejmowałaby:

1. Naukę fizycznej ciągłości; nabywanie oczekiwań, w których obiekty nie znikają ani nie pojawiają się ponownie spontanicznie. U ludzi ten punkt widzenia jest prawdopodobnie bardzo silnie wspierany przez intuicje wizualno-przestrzenne na poziomie modalności, w których ciągłość jest domyślna, i to samo powinno dotyczyć AI.

2. Naukę unikalnej korespondencji. Unikalna korespondencja, jako umiejętność umysłowa, ma tendencję do wzmacniania się przez każde wyzwanie zorientowane na cel, w którym użyteczny obiekt nie może znajdować się w dwóch miejscach jednocześnie.

3. Nauka kompletnego mapowania. Kompletność, wraz z symetrią, jest jednym z głównych nacisków poznawczych wdrażanych przez Copycat w jego modelu tworzenia analogii jako operacji percepcyjnej . Dążenie do kompletności oznacza, że wiszące, niezmapowane obiekty odwracają uwagę od postrzeganej „dobroci" mapowania percepcyjnego. Tak więc może istnieć wsparcie na poziomie modalności dla zauważania wiszących, niezmapowanych obiektów w obrazie.

4. Dzięki obecności tych trzech podstawowych koncepcji możliwe jest abstrahowanie koncepcji kompletnego mapowania przy użyciu relacji unikalnej korespondencji, znanej również jako mapowanie jeden do jednego. My, używając zupełnie innej procedury, nazwalibyśmy tę relację tą samą liczbą co („tożsamością numeronu wytworzoną przez liczenie"). 5. Dzięki mapowaniu jeden do jednego, SI może zauważyć, że wszystkie odpowiedzi na zadanie wyzwania są powiązane ze wspólnym prototypem poprzez relację mapowania jeden do jednego. SI może następnie wyabstrahować koncepcję „pięć" używając prototypu jako przykładu i relacji jako testu.

6. Co dalej? Carl Feynman (komunikacja osobista) zauważa w tym momencie, że relacja mapowania jeden do jednego jest przemienna i przechodnia, a zatem definiuje zbiór klas równoważności; te klasy równoważności okazują się liczbami naturalnymi. Na początku używanie „wykrywania klas równoważności" jako metody poznawczej brzmiało jak oszustwo, ale po namyśle trudno zrozumieć, dlaczego ogólna inteligencja nie miałaby zauważyć, kiedy obiekty o wspólnej relacji do prototypu są podobnie powiązane ze sobą. „Klasa równoważności" może być koncepcją matematyczną, która przypadkowo odpowiada mniej więcej (lub nawet dokładnie) własności percepcyjnej.

7. W tym artykule, ze względu na ograniczenia miejsca, nie rozpatrywano kwestii tworzenia pojęcia superklasy liczby.

Inteligencja deliberatywna musi budować złożone koncepcje z prostych koncepcji, w ten sam sposób, w jaki ewolucja buduje detektory cech wysokiego poziomu ponad detektorami cech niskiego poziomu lub buduje organy przy użyciu tkanek, lub buduje myśli nad koncepcjami lub modalnościami. Istnieją ekologie holoniczne w wyuczonej złożoności koncepcji, w ten sam sposób i z mniej więcej tego samego powodu, dla którego istnieje genetycznie określona struktura holoniczna w wykrywaniu cech na poziomie modalności. Kategorie opisują regularności w percepcji i w ten sposób stają się częścią struktury percepcyjnej, w której wykrywane są dalsze regularności. Jeśli programista zainstaluje na stałe subitizer, który wyprowadza numerony (unikalne znaczniki liczbowe) jako wykryte cechy, SI może być w stanie bardzo szybko podzielić „pięć", ale powstały koncept będzie cierpiał z powodu nieprzejrzystości i izolacji. Koncepcja nie będzie miała niższych poziomów organizacji, które umożliwiłyby natywnym zdolnościom poznawczym SI rozmontowanie i ponowne złożenie koncepcji w przydatne nowe kształty; niezdolność SI do rozłożenia koncepcji jest nieprzejrzystością. Koncepcja nie będzie miała otaczającej ekologii podobnych koncepcji i koncepcji wstępnych, takich jak te, które wynikałyby z naturalnego zdobywania wiedzy przez AI. Procesy poznawcze, które wymagają dobrze zaludnionych ekologii koncepcji, nie będą w stanie działać; AI, która ma „trójkąt", ale nie ma „piramidy", ma mniejsze szanse na pomyślną wizualizację „trójkątnej żarówki". To jest izolacja.

Mikrozadania

W modelu DGI rozwoju AI koncepcje są abstrahowane od bazy doświadczalnej; doświadczenia są treścią poznawczą w obrębie modalności sensorycznych; a modalności sensoryczne są ukierunkowane na złożone wirtualne mikrośrodowisko. Posiadanie doświadczeń, z których można abstrahować koncepcję, jest (koniecznym, ale niewystarczającym) wymogiem do nauczenia się koncepcji. W jaki sposób AI uzyskuje te doświadczenia? Możliwe byłoby nauczenie AI „pięćdziesiątki" po prostu poprzez przedstawienie AI serii obrazów sensorycznych (programowo manipulując mikrośrodowiskiem AI) i nakłonienie procesów percepcyjnych AI do ich uogólnienia, ale oddziela to zadanie formowania koncepcji od jej ekologicznej ważności (mówiąc metaforycznie). Cele wiedzy (omówione w późniejszych sekcjach) nie są arbitralne; wynikają z celów świata rzeczywistego lub celów wiedzy wyższego poziomu. Cele wiedzy istnieją w holonicznej ekologii celów; ekologia celów kształtuje nasze cele wiedzy i tym samym często kształtuje samą wiedzę. Pierwszym przybliżeniem ważności ekologicznej jest przedstawienie SI „wyzwania" w jednym z wcześniej zalecanych wirtualnych mikrośrodowisk – na przykład w mikrośrodowisku bilardowym. Odtąd będę skracał „wyzwanie mikrośrodowiskowe" do „mikrozadania". Mikrozadania mogą uczyć pojęć, przedstawiając SI wyzwanie, które musi zostać rozwiązane przy użyciu pojęcia, którego programista chce uczyć. Aby uzyskać skrupulatną ważność ekologiczną, kluczowe pojęcie powinno być częścią większego problemu, ale nawet odgrywanie „jedna z tych rzeczy nie jest taka jak pozostałe" nadal byłoby lepsze niż bezpośrednie manipulowanie procesami percepcyjnymi SI.

Nauczanie pojęcia jako klucza do mikrozadania zapewnia, że podstawowy „kształt" pojęcia i powiązane doświadczenia są tymi wymaganymi do rozwiązania problemów, a SI ma doświadczenie konieczności pojęcia, doświadczenie odkrywania pojęcia i doświadczenie skutecznego używania pojęcia. Efektywna inteligencja powstaje nie poprzez posiadanie pojęć, ale poprzez używanie pojęć; człowiek uczy się używać pojęć, używając ich. AI musi posiadać doświadczenia odkrywania i używania koncepcji, tak jak AI musi posiadać rzeczywiste odniesienia doświadczalne, które koncepcja uogólnia; AI potrzebuje doświadczenia kontekstów, w których koncepcja jest użyteczna. Uformowanie złożonej koncepcji wymaga przyrostowej ścieżki do tej złożonej koncepcji – serii koncepcji bloków konstrukcyjnych i koncepcji prekursorów, tak aby ostatnim krokiem był skok o wielkości, którą można zarządzać. W modelu rozwoju mikrozadań byłoby to wdrażane za pomocą serii mikrozadań o rosnącym stopniu trudności i złożoności, aby nakłonić AI do uformowania koncepcji prekursorów prowadzących do

uformowania koncepcji złożonych i koncepcji abstrakcyjnych. Jest to duży wydatek w wysiłku programisty, ale argumentowałbym, że jest to konieczny wydatek na tworzenie bogatych koncepcji z zorientowanymi na cel podstawami doświadczalnymi. Doświadczalna ścieżka do „piątki" miałaby kulminację w mikrozadaniu, które można by rozwiązać tylko poprzez abstrakcję i użycie koncepcji piątki, i prowadziłaby do tego wyzwania poprzez mikrozadania, które można by rozwiązać tylko poprzez abstrakcję i użycie koncepcji takich jak „ciągłość obiektu", „unikalna korespondencja", „mapowanie", „wiszący członkowie grupy" i przedostatnia koncepcja „mapowania jeden do jednego".

Jeśli chodzi o konkretny protokół mikrozadań, który ma na celu przedstawienie „wyzwania" dla AI, istnieje wiele możliwych strategii. Osobiście wyobrażam sobie prosty protokół mikrozadań (na poziomie „jedna z tych rzeczy nie jest taka jak pozostałe") jako składający się z szeregu „bram", z których każda musi zostać „przekroczona" przez podjęcie jednej z możliwych akcji, w zależności od tego, co AI uważa za regułę wskazującą na poprawną akcję. Przekroczenie dziesięciu kolejnych bramek przy pierwszej próbie jest wskaźnikiem sukcesu. (W przypadku wyboru binarnego szansa, że wydarzy się to przypadkowo, wynosi 1024:1. Jeśli AI myśli wystarczająco szybko, że może się to zdarzyć losowo (co wydaje się raczej mało prawdopodobne), liczba wymaganych kolejnych bramek może zostać zwiększona do dwudziestu lub więcej). W ten sposób AI może odnieść sukces lub ponieść porażkę w przypadku poszczególnych bramek, zbierając dane o poszczególnych przykładach wspólnej reguły, ale nie będzie w stanie wygrać w całym mikrozadaniu, dopóki wspólna reguła nie zostanie pomyślnie sformułowana. Wymaga to mikrośrodowiska zaprogramowanego tak, aby zapewnić nieskończoną (lub po prostu „stosunkowo dużą") liczbę wariantów podstawowego wyzwania – wystarczającą liczbę wariantów, aby uniemożliwić SI rozwiązanie problemu za pomocą prostej pamięci. Wygląd sensoryczny mikrozadania różniłby się w zależności od modalności. W przypadku modalności bilarda Newtona pojedyncza „bramka" (podzadanie) mogłaby składać się z czterech „systemów opcji", przy czym każdy system opcji byłby zgrupowany w „opcję" i „przycisk". Przestrzenne separacje w modalności Newtona byłyby używane do sygnalizowania grupowania; odległość między systemami opcji byłaby duża w stosunku do odległości w obrębie systemów opcji, a odległość między opcją a przyciskiem byłaby duża w stosunku do odległości między podelementami opcji. Każda opcja miałaby inną konfigurację; SI wybrałaby jedną z czterech opcji na podstawie swojej bieżącej hipotezy dotyczącej rządzącej reguły. Na przykład SI mogłaby wybrać opcję składającą się z czterech bilardów lub opcję z dwoma dużymi bilardami i jednym małym bilardem lub opcję z ruchomymi bilardami. Po wybraniu opcji, SI manipulowałaby efektorem silnika bilarda – ucieleśnieniem SI w tym środowisku – w celu dotknięcia przycisku należącego do (zgrupowanego z) wybranej opcji. SI otrzymywałaby następnie sygnał – być może ruch jakiegoś bilarda działającego jako „flaga" – który symbolizowałby sukces lub porażkę. Środowisko przesuwałoby się następnie do następnej „bramy", powodując odpowiednią zmianę w danych sensorycznych do modalności bilarda SI. Ponieważ format mikrozadania jest złożony i wymaga, aby SI zaczynała od zrozumienia pojęć takich jak „przycisk" lub „przycisk należący do wybranej opcji", istnieje oczywisty problem typu „co było pierwsze – jajko czy kura" w nauczaniu SI formatu mikrozadania, zanim mikrozadania będą mogły być używane do nauczania innych koncepcji. Na razie założymy bootstrapping małej bazy koncepcji, być może poprzez „oszukiwanie" i używanie treści poznawczych stworzonych przez programistów jako tymczasowego rusztowania. Biorąc pod uwagę ten format wyzwania, proste mikrozadanie dla „piątkowatości" wydaje się proste: opcja zawierająca pięć bilardów, niezależnie od ich rozmiaru, względnych pozycji lub wzorców ruchu, jest kluczem do bramki. W praktyce skonfigurowanie mikrozadania dla „piątkowatości" może okazać się trudniejsze ze względu na konieczność wyeliminowania różnych fałszywych sposobów dochodzenia do rozwiązania. W szczególności, jeśli SI ma wystarczająco szeroką gamę detektorów cech ilościowych, to SI będzie prawie na pewno posiadać wyłaniający się Model Akumulatora liczenia. Jeśli SI potrzebuje stosunkowo stałej ilości czasu na mentalne przetworzenie każdego obiektu, to klastrowanie pojedynczych cech na

subiektywnie postrzeganym czasie mentalnego przetworzenia grupy mogłoby dać rozwiązanie mikrozadania bez złożonej koncepcji „piątkowatości". Zamiast „piątkowatości" SI utworzyłaby koncepcję „rzeczy-których-zrozumienie-zajmuje-około-20-milisekund". Rzeczywisty odpowiednik tej sytuacji miał już miejsce, gdy eksperyment, który wcześniej uważano za dowód na niemowlęcą umiejętność liczenia na małych zestawach wizualnych, wykazał wrażliwość na długość konturu (obwód) zestawu wizualnego, ale nie na kardynalność zestawu wizualnego. Nawet jeśli wszystkie koncepcje prekursorowe są już obecne, złożone mikrozadanie może być konieczne, aby uczynić pięciokąt najprostszą poprawną odpowiedzią. Ponadto mikrozadania dla wcześniejszych koncepcji prowadzących do pięciokątności mogą z natury wymagać większej złożoności niż protokół „zestawu opcji" opisany powyżej. Koncepcja unikalnej korespondencji wywodzi swoje zachowanie z właściwości fizycznych. Wybór właściwego zestawu opcji jest zadaniem decyzyjnym percepcyjnym, a nie zadaniem manipulacji fizycznej; w mikrozadaniu decyzyjnym jedynym manipulacyjnym podzadaniem jest manewrowanie bilardem efektorowym w celu dotknięcia wybranego przycisku. Koncepcje takie jak „wiszące obiekty" lub „mapowanie jeden do jednego" mogą wymagać podzadań manipulacyjnych, a nie podzadań decyzyjnych, aby włączyć do koncepcji informacje zwrotne o wynikach fizycznych (mikrośrodowiskowych). Na przykład mikrozadanie do nauczania „mapowania jeden do jednego" może obejmować odpowiednik problemu kołka i dziurki w mikroświatach. Mikrozadanie może polegać na podzieleniu 9 „kołków" na 9 „otworów" – gdzie 9 „otworów" jest podzielonych na trzy podgrupy po 4, 3 i 2, a SI musi z góry przydzielić zapas kołków tym podgrupom. Na przykład w pierwszym etapie mikrozadania SI może mieć pozwolenie na przenoszenie kołków między trzema „pokojami", ale nie może umieszczać kołków w otworach. W drugim etapie mikrozadania SI będzie próbowała umieścić kołki w otworach i odniesie sukces lub porażkę w zależności od tego, czy początkowy przydział między pokojami był prawidłowy. Ze względu na złożoność tego mikrozadania, może ono wymagać innych mikrozadań, aby po prostu wyjaśnić format problemu – nauczyć AI o kołkach, otworach i pokojach. („Kołki i otwory" są uniwersalne i łatwo można je przełożyć na modalność bilarda; „otwory" na przykład mogą być nieruchomymi bilardami, a „kołki" ruchomymi bilardami, które mają być umieszczone w kontakcie z „otworami"). Umieszczanie wirtualnych kołków w wirtualnych otworach nie jest z natury imponującym wynikiem. W tym przypadku AI uczy się rozwiązywania prostego problemu, tak aby wyuczona złożoność przeniosła się na rozwiązywanie złożonych problemów. Jeśli wyuczona złożoność zostanie przeniesiona, a AI później będzie rozwiązywać trudniejsze wyzwania, to z perspektywy czasu sprawienie, aby AI myślała wystarczająco spójnie, aby poruszać się po mikrozadaniu, „będzie" imponującym wynikiem.

Interakcje na poziomie koncepcji

Interakcje koncepcja-koncept są łatwiej dostępne dla introspekcji i technik eksperymentalnych i są stosunkowo dobrze znane w AI i psychologii poznawczej. Podsumowując złożoność interakcji między pojęciami:

• Koncepcje są powiązane z innymi koncepcjami. Aktywowanie koncepcji może „przygotowywać" pobliską koncepcję, gdzie „przygotowywanie" jest zwykle mierzone eksperymentalnie w kategoriach skróconych czasów reakcji. Sugeruje to, że więcej zasobów obliczeniowych powinno być przeznaczonych na skanowanie przygotowanych koncepcji lub że przygotowane koncepcje powinny być najpierw skanowane. (Ten punkt widzenia jest zbyt mechanomorficzny, aby można go było uznać za wyjaśnienie przygotowywania u ludzi. Preaktywacja lub wcześniejsze wiązanie sieci neuronowej byłoby bardziej realistyczne.)

• Pobliskie koncepcje mogą czasami „przesuwać się" pod presją poznawczą; na przykład „trójkąt" do „piramidy". Takie przesunięcia odgrywają główną rolę w analogiach w ramach systemu Copycat Przesunięcia występujące w złożonych problemach projektowania i planowania prawdopodobnie

obejmują wrażliwość na kontekst, a nawet orientację na cel; zobacz późniejszą dyskusję na temat konfliktu i rezonansu w obrazowaniu mentalnym.

• Koncepcje, w swojej roli jako kategorie, dzielą terytorium. Pojedynczy wróbel, jako obiekt, jest opisywany przez koncepcje „wróbel" i „ptak". Wszystkie obiekty, które można opisać jako „wróbel", będą również opisane przez „ptaka". Tak więc informacje docierające przez „ptaka" zazwyczaj, choć nie zawsze, będą dotyczyć całego terytorium „wróbla". Ta forma dziedziczenia może mieć miejsce bez wyraźnej reguły „jest-czymś" łączącej „wróbla" z „ptakiem"; wystarczy, że „ptak" opisuje wszystkie desygnaty „wróbla".

• Koncepcje, w swojej roli jako kategorie, mają relacje superkategorii i podkategorii. Deklaratywne przekonania ukierunkowane na koncepcje mogą być czasami dziedziczone przez takie powiązania. Na przykład „Co najmniej jedno X jest A" jest dziedziczone przez superkategorię Y X: Jeśli wszystkie desygnaty X są desygnatami Y, to „Co najmniej jedno desygnaty X jest A" oznacza, że „Co najmniej jedno desygnaty Y jest A". Odwrotnie, reguły takie jak „Wszystkie X są A" są dziedziczone przez podkategorie X, ale nie przez superkategorie X. Dziedziczenie, które ma miejsce na poziomie koncepcji, poprzez regułę „jest-a", należy odróżnić od pseudodziedziczenia, które ma miejsce poprzez wspólne terytorium w określonych obrazach mentalnych. Kwantyfikatory mentalne, takie jak „wszystkie X są Y", zwykle tłumaczą się na „większość X jest Y" lub „X, domyślnie, jest Y"; wszystkie przekonania podlegają kontrolowanemu wyjątkowi. Możliwe jest rozumowanie o hierarchiach kategorii w sposób rozmyślny, a nie percepcyjny, ale nasza szybkość w tym sugeruje percepcyjny skrót.

• Koncepcje posiadają relacje transformacji, które są ponownie zilustrowane w Copycat. Na przykład w Copycat „a" jest „poprzednikiem" „b", a „1" jest „poprzednikiem" „2". W ogólnej inteligencji te relacje koncepcja-pojęcie odnosiłyby się do obserwacji procesów transformacyjnych działających na referencje doświadczalne, które powodują, że ten sam ciągły obiekt przesuwa się z jednej kategorii do drugiej, i byłyby z nich uogólniane. Często kategorie powiązane przez procesy transformacyjne są podkategoriami tej samej superkategorii.

• Pojęcia działają jako czasowniki, przymiotniki i przysłówki, a także rzeczowniki. U ludzi pojęcia działają jako predykaty jedno-, dwu- i trzymiejscowe, co ilustrują „podmiot", „dopełnienie bliższe" i „dopełnienie pośrednie" w ludzkich częściach mowy; „X daje Y Z". U ludzi predykaty czteromiejscowe i wyższe są prawdopodobnie reprezentowane przez reguły proceduralne, a nie percepcyjnie; spontaniczne zauważenie predykatu czteromiejscowego może być bardzo kosztowne obliczeniowo. Odkrycie relacji predykatu jest wspomagane przez kategoryzację podmiotów predykatu, wykluczając złożoność nieistotną dla predykatu.

• Koncepcje, w swojej roli symboli z tagami słuchowymi, wizualnymi lub gestykulacyjnymi, odgrywają fundamentalną rolę zarówno w komunikacji międzyludzkiej, jak i wewnętrznej ludzkiej konceptualizacji. Krótki, chwytliwy tag słuchowy „pięć" może reprezentować złożoność związaną z koncepcją piątki. Dwoje ludzi, którzy mają wspólną bazę leksykalną, może komunikować złożony obraz mentalny, interpretując obraz za pomocą pojęć, opisując obraz za pomocą struktury pojęć, tłumacząc pojęcia w ramach struktury na społecznie współdzielone tagi słuchowe, przekształcając strukturę pojęć w sekwencję liniową za pomocą wspólnej składni i emitując tagi słuchowe w tej liniowej sekwencji. (Aby przetłumaczyć poprzednie zdanie na język angielski: Komunikujemy się za pomocą zdań, które używają słów i składni ze wspólnego języka.) Ta sama podstawa złożoności jest najwyraźniej również używana do podsumowywania i zwartego manipulowania myślami wewnętrznie; zobacz następną sekcję.

Poziom myśli

Pojęcia są kombinatoryczną wyuczoną złożonością. Pojęcia reprezentują regularności, które powtarzają się nie w izolacji, ale w połączeniu i interakcji z innymi takimi regularnościami. Regularności nie są izolowane i niezależne, ale są podobne do innych regularności, a istnieją prostsze regularności i bardziej złożone regularności, tworząc metaforyczną „ekologię" regularności. Ten zasadniczy fakt dotyczący struktury naszego wszechświata o niskiej entropii sprawia, że inteligencja jest możliwa, możliwa do obliczenia, możliwa do rozwinięcia w obrębie genotypu i możliwa do nauczenia w obrębie fenotypu. Poziom myśli leży ponad wyuczoną złożonością poziomu pojęć. Myśli są strukturami kombinatorycznych pojęć, które zmieniają obrazowanie w przestrzeni roboczej modalności sensorycznych. Myśli są jednorazowymi, jednorazowymi strukturami wdrażającymi nierekurencyjny umysł w nierekurencyjnym świecie. Modalności są okablowane; pojęcia są wyuczone; myśli są wymyślane. Tam, gdzie pojęcia są elementami konstrukcyjnymi, myśli są natychmiastowe. Czasami odległość między pojęciem a myślą jest bardzo krótka; ptak jest koncepcją, ale przy niewielkim wysiłku może stać się myślą, która przywołuje przykład ptaka jako konkretne wyobrażenie mentalne. Niemniej jednak nadal istnieje różnica koncepcyjna między cegłą a domem, który został zbudowany z jednej cegły. Koncepcje, uważane za koncepcje, są elementami konstrukcyjnymi z gotowymi do użycia jądrami koncepcji. Myśl wypełnia wszystkie luki i tłumaczy kombinatoryczne koncepcje na konkretne wyobrażenia mentalne, nawet jeśli myśl jest zbudowana z pojedynczego konceptu. Koncepcje znajdują się w długoterminowym magazynie; myśli wpływają na konkretne wyobrażenia. Widma dla „nauczonego kontra wymyślonego", „kombinatoryjnego kontra specyficznego", „przechowywanego kontra instancjonowanego" i „powracającego kontra nierekurencyjnego" są koncepcyjnie oddzielne, chociaż głęboko ze sobą powiązane i zwykle skorelowane. Pewna treść poznawcza rozciąga się na poziomach koncepcji i myśli. „Przekonania" (wiedza deklaratywna) są wyuczone, specyficzne, przechowywane i powtarzające się. Pamięć epizodyczna w magazynie jest wyuczona, specyficzna, przechowywana i nierekurencyjna. Możliwe są jeszcze drobniejsze gradacje: odzyskana pamięć epizodyczna jest wyuczona, konkretna i natychmiastowa; pamięć może powracać jako treść mentalna, ale jej zewnętrzny referent jest nierekurencyjny. Podobnie, koncepcja, która odnosi się do konkretnego zewnętrznego obiektu, jest wyuczona, konkretna, przechowywana i „półrekurencyjna" w tym sensie, że może odnosić się do więcej niż jednego obrazu sensorycznego, ponieważ obiekt może być napotkany więcej niż raz, ale nadal odnosi się tylko do jednego obiektu, a nie do ogólnej kategorii.

Myśli i język

Archetypowymi przykładami „myśli" (wymyślonych, konkretnych, urzeczywistnionych, nierekurencyjnych) są zdania mentalnie „wypowiedziane" i mentalnie „słyszane" w ludzkim strumieniu świadomości. Używamy tego samego rodzaju zdań, wypowiadanych na głos, aby komunikować myśli między ludźmi. Słowa to znaczniki fonemiczne (mowa), znaczniki wizualne (pisanie), znaczniki gestów (język migowy) lub znaczniki dotykowe (Braille) używane do przywoływania pojęć. Odtąd będę używać mowy do oznaczania wszystkich modalności językowych; „znacznik słuchowy" lub „znacznik fonemiczny" należy rozumieć jako oznaczający znacznik w dowolnej modalności. Gdy mniej więcej ten sam koncept dzieli mniej więcej ten sam znacznik fonemiczny w grupie ludzi, słowa mogą być używane do komunikowania pojęć między ludźmi, a zdania mogą być używane do komunikowania złożonych obrazów. Fonemy słowa mogą wywoływać całą funkcjonalność rzeczywistego konceptu związanego ze znacznikiem słuchowym. Zdanie mówione jest liniową sekwencją słów; Mózg człowieka używa reguł gramatycznych i składniowych, aby złożyć liniową sekwencję w strukturę pojęć, kompletną z wewnętrznymi i zewnętrznymi informacjami docelowymi. „Trójkątna żarówka", przymiotnik, po którym następuje rzeczownik, staje się „trójkątny" celując w „żarówkę". „To jest telefon", anafora-czasownik-artykuł-rzeczownik, staje się stwierdzeniem o telefoniczności wcześniej odniesionego

obiektu. „To” jest odniesieniem wstecznym do wcześniej przywołanego celu mentalnego, więc towarzyszący opis poznawczy („to jest telefon”) jest narzucany obrazom poznawczym reprezentującym odniesienie do „tego”. Proces poznawczy, który buduje strukturę pojęcia z sekwencji słów, łączy ograniczenia składniowe i semantyczne; czysta składnia jest szybsza i wyprzedza semantykę, ale dysharmonie semantyczne mogą rozbić syntaktycznie wytworzone struktury poznawcze. Semantyczne przewodniki do interpretacji sięgają również poziomu słowa, wpływając na interpretację homofonów i niejednoznacznych fonemów. Na razie pozostawię otwarte pytanie, dlaczego słyszymy „zdania mentalne” wewnętrznie – to znaczy, dlaczego transformacja struktur pojęciowych w liniowe sekwencje słów, oczywiście niezbędna do komunikacji mówionej, zachodzi również wewnętrznie w strumieniu świadomości. Później spróbuję wyjaśnić to jako wynikające z koewolucji myśli i języka. Na razie niech pozostanie fakt, że kombinatoryczna struktura słów i zdań w naszej wewnętrznej narracji odzwierciedla kombinatoryczną strukturę pojęć i myśli.

Obrazowanie mentalne

Złożoność poziomu organizacji myśli wynika z cyklicznej interakcji myśli i obrazów mentalnych. Myśli modyfikują obrazy mentalne, a obrazy mentalne z kolei dają początek myślom. Obrazy mentalne istnieją w reprezentacyjnej przestrzeni roboczej modalności sensorycznych. Obrazy sensoryczne wynikają z informacji środowiskowych (niezależnie od tego, czy środowisko jest „rzeczywiste” czy „wirtualne”); obrazy wyobrażeniowe wynikają z manipulacji przestrzenią roboczą modalności poprzez narzucanie pojęć i wyszukiwanie w pamięci. Obrazy mentalne, czy to sensoryczne czy wyobrażeniowe, wykazują organizację holoniczną: z poziomu „pikseli” do obiektów i fragmentów; z obiektów i fragmentów do grup i superobiektów; z grup i superobiektów do scen mentalnych. W ludzkim widzeniu przykładami konkretnych zasad rządzących grupowaniem są bliskość, podobieństwo koloru, podobieństwo rozmiaru, wspólny los i zamknięcie; kontynuacja; wspólny region i łączność; i współliniowość. Niektóre z zaproponowanych paradygmatów, które zostały zaproponowane w celu rozwiązania pozytywnych danych wejściowych z zasad grupowania i negatywnych danych wejściowych z wykrytych konfliktów, w spójną globalną organizację, obejmują: holoniczne rozwiązywanie konfliktów (opisane wcześniej), temperaturę obliczeniową , Prägnanz , sieci Hopfielda , zasadę prawdopodobieństwa , minimalną długość opisu i propagację ograniczeń. Obrazy mentalne zapewniają przestrzeń roboczą dla określonych percepcji pojęć i struktur pojęć. Fragment obrazów sensorycznych może zostać mentalnie oznaczony strukturą pojęć „żółte pole”, a opis ten pozostanie związany z obiektem – częścią percepcji obiektu – nawet poza zakresem bezpośredniej myśli. Wyuczone kategorie i wyuczone oczekiwania wpływają również na organizację gestalt obrazów mentalnych . Obrazy mentalne są aktywnym płótnem, na którym malowana jest myśl deliberatywna – „aktywne płótno” oznaczające dynamiczny proces, a nie tylko statyczną reprezentację. Gestalt obrazowania mentalnego jest produktem wielu lokalnych relacji między elementami. Ponieważ automatyczne procesy poznawcze utrzymują gestalt, lokalna zmiana w obrazowaniu może mieć konsekwencje dla połączonych elementów w obrazowaniu roboczym, bez konieczności określania tych zmian w bliskiej myśli, która spowodowała modyfikację. Spójność gestalt obrazowania zapewnia również informacje zwrotne na temat tego, które możliwe zmiany będą dobrze spójne, i dlatego jest jednym z czynników weryfikujących, które potencjalne myśli osiągną status aktualności (patrz poniżej). Obrazowanie wspiera abstrakcyjne percepcje. Człowiek może rozumować o obiekcie, o którym wiadomo, że kosztuje 1000 dolarów, ale dla którego nie ma innych informacji mentalnych. Abstrakcyjne rozumowanie o tym obiekcie wymaga sposobu reprezentowania obiektów mentalnych, które nie zajmują żadnej a priori modalności; nie oznacza to jednak, że abstrakcyjne rozumowanie działa niezależnie od wszystkich modalności. Abstrakcyjne rozumowanie może działać poprzez „śledzący obiekty” poziom modalności, który może działać niezależnie od modalności, które śledzi; lub poprzez pożyczenie istniejącej modalności za pomocą metafory (patrz poniżej); lub pierwsza opcja

może być używana rutynowo, a druga opcja, gdy jest to konieczne. Biorąc pod uwagę abstrakcyjny „obiekt, który kosztuje 1000 dolarów", można dołączyć struktury pojęciowe, które opisują obiekt bez konieczności opisywania konkretnych obrazów sensorycznych. Jeśli narzucę koncepcję „czerwony" na istniejący abstrakcyjny obraz dla „obiektu, który kosztuje 1000 dolarów", aby uzyskać „czerwony obiekt, który kosztuje 1000 dolarów", koncepcja „czerwony" wisi tam, gotowa do aktywacji, gdy tylko będzie mogła, ale jeszcze nie dająca konkretnych obrazów wizualnych. Podobnie, wiedza uogólniona z doświadczenia w relacjach koncepcja-pojęcie może być użyta do wykrywania abstrakcyjnych konfliktów. Jeśli wiem, że wszystkie pingwiny są zielone, mogę wywnioskować, że „czerwony obiekt, który kosztuje 1000 dolarów" nie jest pingwinem. Możliwe jest wykrycie konfliktu między „czerwonym" i „zielonym" poprzez porównanie na poziomie koncepcji dwóch abstrakcyjnych opisów, nawet w przypadku braku wizualizowanych obrazów mentalnych. Nie oznacza to jednak, że rozwój AI może implementować wyłącznie „abstrakcyjne rozumowanie" i pomijać modalności sensoryczne. Po pierwsze, prawdziwy umysł wykorzystuje bogatą złożoność na poziomie pojęć uzyskaną z doświadczenia sensorycznego i z doświadczenia rozumowania, które wykorzystuje w pełni zwizualizowane wyobrażeniowe obrazy, aby wspierać abstrakcyjne rozumowanie; wiemy, że „czerwony" kłóci się z „zielonym" z powodu wcześniejszego doświadczenia sensorycznego z czerwonym i zielonym. Po drugie, samo to, że niektóre kroki w rozumowaniu wydają się teoretycznie możliwe do przeprowadzenia wyłącznie na poziomie pojęć, nie oznacza, że kompletny proces deliberatywny może być przeprowadzony wyłącznie na poziomie pojęć. Po trzecie, abstrakcyjne rozumowanie często wykorzystuje metaforę, aby dodać zachowania modalności do procesu abstrakcyjnego rozumowania. Idea „czystego" abstrakcyjnego rozumowania historycznie doprowadziła do patologii AI i powinna być uważana za szkodliwą. Mając na uwadze tę ostrożność, jest jednak możliwe, że umysły ludzkie wizualizują koncepcje tylko w zakresie wymaganym przez obecny tok myślenia, oszczędzając w ten sposób zasoby umysłowe. Wczesna AI prawdopodobnie będzie mniej biegła w tej sztuczce, co oznacza, że wczesne AI mogą potrzebować używać pełnych wizualizacji, podczas gdy człowiek mógłby użyć abstrakcyjnego rozumowania. Abstrakcyjne rozumowanie jest środkiem, za pomocą którego indukcyjnie nabyte uogólnienia mogą być używane w rozumowaniu dedukcyjnym. Jeśli empiryczna indukcja z bazy doświadczalnej, w której wszystkie obserwowane pingwiny są zielone, prowadzi do ukształtowania przekonania „pingwiny są zielone", to przekonanie to może odnosić się abstrakcyjnie do „czerwonego obiektu, który kosztuje 1000 dolarów", aby wywnioskować, że ten obiekt prawdopodobnie nie jest pingwinem. W tym przykładzie abstrakcyjne przekonanie jest połączone z abstrakcyjnym obrazowaniem konkretnego obiektu, aby doprowadzić do dalszego abstrakcyjnego wniosku na temat tego konkretnego obiektu. Ludzie wykraczają poza to, stosując bardzo potężną technikę „rozumowania dedukcyjnego". Używamy abstrakcyjnych przekonań, aby rozumować o abstrakcyjnym obrazowaniu mentalnym, które opisuje klasy, a nie tylko konkretne obiekty, i dochodzimy do wniosków, które następnie stają się nowymi abstrakcyjnymi przekonaniami; możemy używać rozumowania dedukcyjnego, jak również rozumowania indukcyjnego, aby nabyć nowe przekonania. „Czyste" rozumowanie dedukcyjne, podobnie jak „czyste" rozumowanie abstrakcyjne, należy uznać za szkodliwe; rozumowanie dedukcyjne jest zwykle ugruntowane w naszej zdolności do wizualizacji konkretnych przypadków testowych i przez przecięcie się potwierdzenia indukcyjnego z wnioskami dedukcyjnymi. Obrazowanie wspiera śledzenie zależności, funkcji poznawczej, która jest koncepcyjnie oddzielona od percepcji przyczynowości zdarzeń. Innym sposobem myślenia o tym jest to, że postrzegana przyczynowość poznawcza nie powinna być mylona z postrzeganą przyczynowością w odniesieniach ze świata rzeczywistego. Mogę wierzyć, że słońce wkrótce wzejdzie; przyczyną tego przekonania może być to, że usłyszałem pianie koguta; mogę wiedzieć, że moja pewność co do bliskości wschodu słońca opiera się na mojej pewności co do dokładności koguta; ale nie wierzę, że pianie koguta powoduje wschód słońca. Obrazowanie wspiera złożone percepcje „pewności" poprzez śledzenie polegania na źródłach niepewności. Biorąc pod uwagę

twierdzenie A z 50% pewnością, że „obiekt X jest niebieski” i przekonanie B z 50% pewnością, że „niebieskie obiekty są duże”, klasycznym wnioskiem byłoby twierdzenie „obiekt X jest duży” z 25% pewnością. Jednak ta prosta metoda arytmetyczna pomija możliwość, ważną nawet w logice klasycznej, że A i B są wzajemnie zależne od trzeciej niepewności C – w takim przypadku łączna pewność może być większa niż 25%. Na przykład, w przypadku gdy „obiekt X jest niebieski” i „niebieskie obiekty są duże” są prostymi dedukcjami z trzeciego twierdzenia C z 50% pewnością, a ani A, ani B nie mają żadnej własnej wrodzonej niepewności, to „obiekt X jest duży” jest również prostym dedukcją z C i ma pewność 50%, a nie 25%. Pewności nie należy postrzegać jako pojedynczego prawdopodobieństwa ilościowego; pewność jest percepcją, która podsumowuje sieć polegania na źródłach niepewności. Proste powiązania – to znaczy powiązania, których lokalna niepewność jest tak niska, że jest nieistotna – mogą zostać wyeliminowane z postrzeganych zależności dedukcji w przód: „obiekt X jest duży” jest postrzegane jako dedukcja twierdzenia C, a nie dedukcja z C plus „obiekt X jest niebieski” plus „niebieskie obiekty są duże”. Jeśli jednak twierdzenie „obiekt X jest niebieski” jest sprzeczne z niezależnymi dowodami wspierającymi niespójne twierdzenie „obiekt X jest czerwony”, to poleganie na „obiekcie X jest niebieski” jest niezależnym źródłem niepewności, wykraczającym poza wyprowadzone poleganie na C. Oznacza to, że pewność twierdzenia można ocenić, ważąc je w stosunku do wsparcia dla negacji twierdzenia [101]. Chociaż globalna struktura polegań jest strukturą sieci, lokalne postrzeganie pewności jest bardziej prawdopodobne, że pochodzi z zestawu polegań na wspierających i zaprzeczających twierdzeniach, których niepewność jest wyraźna. To, że lokalne postrzeganie pewności jest zestawem, a nie workiem lub skierowaną siecią, wyjaśnia eliminację wspólnych polegań w dalszych wyprowadzonych propozycjach i zachowanie globalnej struktury sieciowej. U ludzi postrzeganie pewności wykazuje mniej więcej ilościową siłę, a ta wielkość zachowuje się w pewien sposób jak formalizm matematyczny, który nazywamy „prawdopodobieństwem”. Pewność i prawdopodobieństwo nie są identyczne; dla ludzi jest to zarówno zaleta, jak i wada. Postrzeganie twierdzenia opartego na czterech niezależnych twierdzeniach o 80% pewności jako psychologicznie odmiennego od twierdzenia opartego na jednym twierdzeniu o 40% pewności może przyczyniać się do użytecznej inteligencji. Z drugiej strony, ludzka niezdolność do użycia arytmetycznie precyzyjnego traktowania prawdopodobieństw może przyczyniać się do znanych przypadków rozumowania nienormatywnego, takich jak nieuwzględnianie bayesowskich priorów, przecenianie prawdopodobieństw koniunkcyjnych i niedocenianie prawdopodobieństw rozłącznych oraz innych klasycznych błędów . AI najlepiej mogłaby zacząć od oddzielnych perceptów dla pewności „ludzkiej” i pewności „arytmetycznej”. Obrazowanie oddziałuje z informacjami sensorycznymi o swoim referencie. Obrazowanie oczekiwane jest potwierdzane lub naruszane przez rzeczywiste zdarzenie. Abstrakcyjne obrazowanie stworzone i pozostawione wiszące wiąże się z perceptem sensorycznym swojego referenta, gdy i jeśli percept sensoryczny stanie się dostępny. Obrazowanie oddziałuje z bayesowskimi informacjami o swoim referencie: twierdzenia, które formułują przewidywania dotyczące przyszłych informacji sensorycznych, są potwierdzane lub odrzucane, gdy informacje sensoryczne docierają, aby spełnić lub zaprzeczyć przewidywaniu. Potwierdzenie lub zaprzeczenie przekonania może rozprzestrzeniać się wstecz, aby działać jako bayesowskie potwierdzenie lub zaprzeczenie źródeł jego wsparcia. (W takich przypadkach rozumowanie normatywne jest zazwyczaj rządzone przez bayesowskie twierdzenie prawdopodobieństwa). Zdolność wyobrażeń do wiązania się z ich odniesieniem jest określana przez zdolność „dopasowania” wyobrażeń – ich zdolność do odróżniania percepcji sensorycznej jako należącej do nich samych – co z kolei jest właściwością sposobu, w jaki abstrakcyjne wyobrażenia oddziałują z przychodzącymi wyobrażeniami sensorycznymi na aktywnym płótnie pamięci roboczej. Klasyczna sztuczna inteligencja z symbolem „hamburgera” może być w stanie odróżnić poprawnie napisane naciśnięcia klawiszy podczas pisania „hamburgera”, ale nie ma zdolności dopasowywania, aby wiązać się z hamburgerami w jakikolwiek inny sposób, na przykład wizualnie lub węchowo. U ludzi abstrakcyjne wyobrażenia „czerwonego obiektu” mogą nie obejmować konkretnego

czerwonego obrazu, ale koncepcja „czerwonego” jest nadal związana z abstrakcyjnym obrazowaniem, a abstrakcyjne wyobrażenia mogą wykorzystywać jądro „czerwonego”, aby dopasować odniesienie w obrazowaniu sensorycznym. Obrazowanie może wiązać się ze swoim odniesieniem na różne sposoby. Obraz mentalny może być bezpośrednim, środowiskowym doświadczeniem sensorycznym; może być przywołanym wspomnieniem; może być przewidywaniem przyszłych zdarzeń; może odnosić się do teraźniejszości lub przeszłości świata; może być scenariuszem łączącym lub kontrfaktycznym. Możemy oddzielić scenariusz łączący od sceny opisowej, myśląc „Co jeśli?” i ekstrapolując, i możemy oddzielić oddzielny scenariusz łączący od pierwszego, myśląc ponownie „Co jeśli?”. Ludzie nie mogą kontynuować tego procesu w nieskończoność, ponieważ kończy nam się pamięć krótkotrwała, aby śledzić wszystkie zależności, ale mamy wrodzoną zdolność śledzenia. Należy zauważyć, że obrazy mentalne nie mają nieprzezroczystego znacznika wybranego ze skończonego zbioru „subjunktywnego”, „kontrfaktycznego” itd. Stanowiłoby to nadużycie kodu: bezpośrednie programowanie, jako szczególny przypadek, tego, co powinno wynikać z ogólnych zachowań lub wyłonić się z niższego poziomu organizacji. Twierdzenie w obrazach kontrfaktycznych nie jest koniecznie oznaczone specjalnym znacznikiem „kontrfaktyczny”; raczej „kontrfaktyczny” może być nazwą, którą nadajemy zestawowi wewnętrznie spójnych twierdzeń ze wspólną zależnością od twierdzenia, które jest silnie obalone. Podobnie, prognoza nie jest koniecznie twierdzeniem oznaczonym nieprzezroczystym znacznikiem „przewidywanie”; prognozę lepiej postrzegać jako twierdzenie ze wsparciem dedukcyjnym, którego odniesieniem jest przyszłe zdarzenie lub inne odniesienie, dla którego nie nadeszły jeszcze żadne informacje sensoryczne; obrazy prognozy wiążą się następnie z informacjami sensorycznymi, gdy nadchodzą, umożliwiając wykrycie potwierdzenia lub obalenia. Rozróżnienie między „przewidywaniem”, „kontrfaktycznym” i „scenariuszem subjunktywnym” może wynikać z bardziej ogólnych zachowań dotyczących pewności, polegania i odniesienia. Obrazowanie mentalne wspiera postrzeganie podobieństwa i innych relacji porównawczych, zorganizowanych w złożone mapowania, korespondencje i analogie (przy czym Copycat jest najlepszym istniejącym przykładem implementacji AI). Obrazowanie mentalne wspiera oczekiwania i wykrywanie naruszonych oczekiwań (gdzie „przewidywanie” powyżej odnosi się do produktu rozważań, „oczekiwania” są tworzone przez aplikacje koncepcyjne, zachowania modalności lub interakcje gestalt). Obrazowanie mentalne wspiera obrazowanie czasowe i aktywną wyobraźnię procesów czasowych. Obrazowanie mentalne wspiera opis relacji przyczynowych między zdarzeniami i między twierdzeniami, tworząc złożone sieci przyczynowe, które rozróżniają implikację i bezpośrednią przyczynowość [80]. Obrazy mentalne wspierają wiążącą relację „metafory”, umożliwiając rozszerzone rozumowanie przez analogię, tak aby np. wzrokowo-przestrzenne postrzeganie rozwidlającej się ścieżki mogło zostać wykorzystane do przedstawienia i rozumowania na temat zachowania gałęzi if-then-else, przy czym wnioski wyciągnięte z metafory można (tymczasowo) zastosować do odniesienia. Obrazowanie wspiera adnotację dowolnych obiektów dowolnymi perceptami; jeśli chcę mentalnie oznaczyć mój zegarek jako „X”, to będzie to „X”, a jeśli również oznaczę moje słuchawki i pilota jako „X”, to „X” utworzy nową (choć dowolną) kategorię. Ta podsekcja oczywiście nie była w pełni konstruktywnym opisem obrazowania mentalnego. Raczej był to bardzo krótki opis niektórych głównych właściwości potrzebnych obrazowaniu mentalnemu do wspierania poziomu organizacji myśli. Przepraszam, ale aby napisać teorię ogólnej inteligencji w jednym rozdziale, często konieczne jest skompresowanie ogromnej ilości złożoności w jednym zdaniu i bibliografii.

Pochodzenie myśli

Myśli to zdarzenia poznawcze, które zmieniają wyobrażenia mentalne. Z kolei myśli są tworzone przez procesy, które odnoszą się do wyobrażeń mentalnych, tak że rozważania są wdrażane przez cykliczną interakcję myśli modyfikujących wyobrażenia mentalne, co daje początek dalszym myślom. Nie oznacza to, że poziom rozważań jest „naturalnie wyłaniający się” z myśli. Poziom myśli ma specyficzne

cechy pozwalające na myślenie w akapitach, a nie tylko zdaniach – „ciągi myśli" z wewnętrznym pędem, chociaż nie tak dużym pędem, aby przerwanie było niemożliwe. W dowolnym momencie, z ogromnej przestrzeni możliwych myśli, pojedyncza myśl kończy się „wypowiedzianą" w ramach rozważań. Właściwie „jedna myśl na raz" to po prostu ludzki sposób robienia rzeczy, a wystarczająco zaawansowana sztuczna inteligencja może multipleksować lub wielowątkowo rozważać, ale to nie zmienia podstawowego pytania: Skąd pochodzą myśli? Sugeruję, że najlepiej jest podzielić nasz koncepcyjny pogląd na ten proces na dwie części; po pierwsze, produkcja sugerowanych myśli, a po drugie, selekcja myśli, które wydają się „użyteczne" lub „potencjalnie użyteczne" lub „ważne" lub w inny sposób interesujące. W niektórych przypadkach proces, który wymyśla lub sugeruje myśli, może wykonać większość pracy, a selekcja jest stosunkowo nieistotna; gdy przypadkowo położysz rękę na gorącym piecu, wynikające z tego zdarzenie oddolne natychmiast przejmuje rozważania. W innych przypadkach proces selekcji może obejmować większość użytecznej inteligencji, a duża liczba możliwych myśli jest testowana równolegle. Oprócz tego, że jest użyteczny koncepcyjnie, rozróżnianie między sugestią a weryfikacją jest przydatne na poziomie projektowania, jeśli „weryfikatorzy" i „sugerujący" mogą skorzystać z organizacji modułowej. Wielu sugerujących może zostać ocenionych przez jednego weryfikatora, a wielu weryfikatorów może podsumować dobroć sugestii. Nie oznacza to koniecznie sztywnych etapów przetwarzania, w których „sugestia" jest uruchamiana, kończy się i jest ściśle przestrzegana przez „weryfikację", ale oznacza to wspólny grunt, na którym repertuary procesów sugestii i procesów weryfikacji oddziałują na siebie. Używam terminu sequitur, aby odnieść się do procesu poznawczego, który sugeruje myśli. „Sequitur" odnosi się nie do sposobu, w jaki dwie myśli następują po sobie – to jest sfera rozważań – ale raczej do źródła, z którego powstaje pojedyncza myśl, wynikająca z obrazów mentalnych. Nawet zanim sugerowana myśl wypłynie na powierzchnię, sugestia może oddziaływać z obrazami mentalnymi, aby określić, czy myśl jest interesująca i ewentualnie wpłynąć na ostateczną formę myśli. Określone interakcje nazywam rezonansami; sugerowana myśl rezonuje z obrazami mentalnymi podczas weryfikacji. Zarówno pozytywne rezonanse, jak i negatywne rezonanse (konflikty) mogą uczynić myśl bardziej interesującą, ale myśl bez żadnych rezonansów prawdopodobnie nie będzie interesująca. Przykładem sequitur może być zauważenie, że część obrazów mentalnych spełnia koncepcję; dla człowieka oznaczałoby to myśl „X jest Y!". W tym przykładzie koncepcja jest wskazywana i spełniana przez ciągły proces tła, a nie jest sugerowana przez rozważania odgórne; tak więc zauważenie, że X jest Y jest zaskoczeniem, które może zmienić obecny tok myślenia. Jak wielkie zaskoczenie – jak wyraziste stanie się odkrycie – będzie zależało od szeregu otaczających czynników, z których większość to prawdopodobnie te same rezonanse, które promowały kandydaturę sugestii „koncept Y pasuje do X" do prawdziwej myśli „X jest Y!". (Różnica między sugestią a myślą polega na tym, że prawdziwa myśl uporczywie zmienia obecne obrazy mentalne, wiążąc koncepcję Y z X i przesuwając punkt ciężkości uwagi). Jakie są czynniki, które determinują rezonans sugestii „koncept Y pasuje do X" lub „koncept Y może pasować do X" i wyrazistość myśli „X jest Y"? Niektóre z tych czynników będą inherentnymi właściwościami koncepcji Y, takimi jak przeszła wartość Y, rzadkość Y, złożoność Y itd.; w AI są to już znane metody określania względnej wartości heurystyk i względnej wyrazistości kategorii. Inne czynniki są nieodłączne dla X, takie jak stopień, w jakim X jest przedmiotem uwagi. Bardziej skomplikowane czynniki wyłaniają się z interakcji X (docelowy obraz), Y (przechowywany koncept, który potencjalnie pasuje do X), sugerowanego obrazu mentalnego dla Y opisującego X, otaczającego obrazu i kontekstu zadania. Ludzki programista badający ten problem projektowy naturalnie widzi nieograniczony zakres potencjalnych korelacji. Aby uniknąć paniki, należy pamiętać, że ewolucja nie rozpoczęła się od rozważania całej przestrzeni wyszukiwania i próby jej ograniczenia; ewolucja stopniowo rozwinęłaby repertuar korelacji, w którym odpowiednie myśli rezonowałyby przez pewien czas. Tak jak jądra koncepcji nie są kompletne pod względem AI, tak też sekwencje i rezonanse nie są kompletne pod względem AI. Sequitury i rezonanse mogą również nie musieć być równoważne ludziom, aby w minimalnym stopniu wspierać rozważania; dopuszczalne jest,

aby wczesna SI pomijała wiele oczywistych myśli ludzkich, pod warunkiem, że te myśli, które są pomyślnie generowane, sumują się do w pełni ogólnych rozważań. Konkretne sequitury i rezonanse często wydają się przypominać ogólne heurystyki w Eurisko Lenata [60] lub innych programach AI przeznaczonych do wyszukiwania interesujących pojęć i przypuszczeń. Podobieństwo jest jeszcze bardziej wzmocnione przez pomysł dodania do mieszanki wyuczonych skojarzeń; na przykład korelowanie, które pojęcia Y są często przydatne w przypadku obrazów opisanych przez pojęcia X lub korelowanie pojęć uznanych za przydatne w odniesieniu do kategoryzacji bieżącej domeny zadania, przypomina nieco próbę Eurisko nauczenia się konkretnych heurystyk dotyczących tego, kiedy konkretne pojęcia są przydatne. Podobnie, ogólna sekwencja, która przeszukuje skojarzone pojęcia, aby dopasować je do działających obrazów, przypomina nieco Eurisko stosujące heurystykę. Pomimo podobieństwa strukturalnego, sequitury nie są heurystykami. Sequitury to ogólne podprocesy poznawcze leżące na poziomie organizacji oprogramowania mózgowego. Podprocesem jest sequitur, który obsługuje myśli o ogólnej formie „X jest Y"; każda treść poznawcza odnosząca się do konkretnych X i Y jest wyuczoną złożonością, niezależnie od tego, czy przyjmuje formę przekonań heurystycznych, czy też skojarzeń korelatywnych. Ponieważ nasza wewnętrzna narracja jest otwarta na introspekcję, nie dziwi fakt, że sequitury wytwarzają pewne myśli przypominające zastosowanie heurystyk; zdania mentalne wytwarzane przez sequitury są otwarte na introspekcję, a badacze AI przyglądali się tym zdaniom mentalnym, gdy wynaleziono heurystyki. Niektóre myśli, które mogą wynikać z „X jest Y!" (nieoczekiwane zadowolenie z koncepcji) to: „Dlaczego X jest Y?" (poszukiwanie wyjaśnienia); lub „Z oznacza, że X nie może być Y!" (wykrywanie naruszenia przekonania); lub „X nie jest Y" (ponowne sprawdzenie wstępnego wniosku). Każda sekwencja dwóch lub więcej myśli jest technicznie domeną rozważań, ale powiązane rozważania są wspierane przez właściwości poziomu myśli, takie jak skupienie uwagi. Powodem, dla którego „Dlaczego X jest Y?" prawdopodobnie wynika z „X jest Y!" jest to, że myśl „X jest Y" przesuwa skupienie uwagi na Y-owość X (obrazowanie mentalne koncepcji Y wiążącej się z X), tak że procesy sequitur mają tendencję do selektywnego skupiania się na tym fragmencie obrazowania mentalnego i próbowania odkrywania myśli, które go obejmują. Wzajemne oddziaływanie myśli i obrazów ma dalsze właściwości, które wspierają rozważania. „Dlaczego X jest Y?" to myśl, która tworzy lub skupia uwagę na pytaniu – magnesie myśli, który przyciąga możliwe odpowiedzi. Obrazowanie pytań jest zarówno podobne, jak i niepodobne do obrazowania celu. (Więcej o celach później; obecnie liczy się to, jak poziom myśli oddziałuje na cele, a intuicyjna definicja celów powinna wystarczyć do tego). Cel w klasycznym sensie można zdefiniować jako abstrakcyjne obrazy, które „chcą być prawdziwe", co wpływa na poznanie poprzez wpływanie na decyzje i działania AI; AI podejmuje decyzje i działania w oparciu o to, czy przewiduje, że te decyzje i działania doprowadzą do odniesienia celu. Pytania wpływają przede wszystkim na to, które myśli się pojawiają, a nie na to, które decyzje są podejmowane. Pytania są złożonością na poziomie myśli, właściwością obrazów mentalnych i nie należy ich mylić z refleksyjnymi celami stwierdzającymi, że dana wiedza jest pożądana; te dwa elementy są ze sobą bardzo silnie powiązane, ale są odrębne koncepcyjnie. Pytanie jest magnesem na myśli, a cel magnesem na działania. Ponieważ myśli błądzące są (mam nadzieję!) mniej niebezpieczne niż działania błądzące, kwestionowanie (zapytanie) może rozprzestrzeniać się w znacznie bardziej niestrukturalny sposób niż celowość (pożądanie). Obrazowanie celu jest abstrakcyjnym obrazowaniem, którego odniesienie jest sprowadzone do zgodności z opisem celu przez działania AI. Obrazowanie pytań jest również obrazowaniem abstrakcyjnym, ponieważ odpowiedź nie jest jeszcze znana, ale obrazowanie pytań ma bardziej otwarte kryterium satysfakcji. Obrazowanie celów ma tendencję do tego, aby jego odniesienie przyjęło określoną wartość; obrazowanie pytań ma tendencję do tego, aby jego odniesienie przyjęło dowolną wartość. Obrazowanie pytań dla „wyniku zdarzenia E" przyciąga wszelkie myśli o wyniku zdarzenia E; jest to agnostyczne pytanie „Jaki, jeśli w ogóle, jest przewidywany wynik zdarzenia E?" Obrazowanie celów dla „wyniku zdarzenia E" ma tendencję do wymagania pewnego określonego wyniku dla E. Tworzenie obrazów pytań jest jednym z głównych

czynników przyczyniających się do ciągłości sekwencji myśli, a zatem jest konieczne do rozważań. Jednak tak jak obrazy celów muszą wpływać na rzeczywiste decyzje i rzeczywiste działania, zanim przyznamy, że AI ma coś, co zasługuje na miano „celu", obrazy pytań muszą wpływać na rzeczywiste myśli – rzeczywiste następstwa i rzeczywiste weryfikatory – aby można je było uznać za poznawczo rzeczywiste pytanie. Jeśli istnieje wyrazisty obraz pytań dla „wyniku zdarzenia E", staje się on celem następstw, które poszukują przekonań o implikacji lub związku przyczynowo-skutkowym, których poprzedniki są spełnione przez aspekty E; innymi słowy, następstwa poszukują przekonań w formie „E zwykle prowadzi do F" lub „E powoduje F". Jeśli istnieje otwarty obraz pytań dla „przyczyny Y -owości X", a myśl sugerowana z jakiegoś innego powodu przypadkowo przecina się z „przyczyną Y -owości X", myśl ta silnie rezonuje i wypływa na powierzchnię poznania. Podobnym i szczególnie znanym następstwem jest poszukiwanie przekonania przyczynowego, którego następnik pasuje do wyobrażeń celu, a którego poprzednik jest następnie wizualizowany jako wyobrażenie opisujące zdarzenie, które ma prowadzić do celu. Stworzone wyobrażenie zdarzenia może stać się nowym wyobrażeniem celu – podcelem – jeśli powiązanie predykcyjne zostanie potwierdzone i nie zostaną oddzielnie przewidziane żadne nieprzyjemne skutki uboczne (więcej informacji o celach i podcelach można znaleźć w dyskusji na temat poziomu deliberacji). Wiele klasycznych teorii AI, w szczególności „dowodzenie twierdzeń" i „planowanie", utrzymuje uproszczoną formę następstwa „poszukiwacza podcelu" jako podstawowego algorytmu ludzkiej myśli. Jednak następstwo to samo w sobie nie wdraża planowania. Proces poszukiwania podcelów jest czymś więcej niż jednym procesem poznawczym poszukiwania następstw przekonań, które pasują do istniejących celów. Istnieją inne drogi do znalezienia kandydatów na podcele poza wstecznym łączeniem istniejących celów; na przykład, rozumowanie do przodu z dostępnych działań. Może istnieć kilka różnych rzeczywistych sekwencji (procesów poznawczych), które poszukują odpowiednich przekonań; podejście projektowe ewolucji polegałoby na „znalezieniu procesów poznawczych, które tworzą użyteczne sugestie", a nie na „ograniczeniu wyczerpującego przeszukiwania wszystkich przekonań, aby uczynić je obliczeniowo wydajnym", co oznacza, że w repertuarze może istnieć kilka sekwencji, które selektywnie poszukują różnych rodzajów przekonań przyczynowych. Znalezienie przekonania, którego następnik pasuje do wyobrażeń celu, nie jest tym samym, co znalezienie zdarzenia, które, jak się przewiduje, doprowadzi do zdarzenia celu; a nawet znalezienie działania, co do którego przewiduje się, że doprowadzi do co najmniej jednego zdarzenia celu, nie jest tym samym, co sprawdzenie czystej pożądaności tego działania. Sequitur, który poszukuje przekonań, których następniki pasują do wyobrażeń celu, jest tylko jednym ze składników poziomu organizacji myśli. Jest to jednak składnik, który wygląda jak „wykrzyknik myśli" z perspektywy wielu tradycyjnych teorii, dlatego warto przejrzeć, w jaki sposób inne poziomy organizacji przyczyniają się do efektywnej inteligencji sequitur „poszukiwacza podcelu". Celem jest opisowy obraz mentalny, prawdopodobnie przyjmujący formę koncepcji lub struktury koncepcji opisującej zdarzenie; myślenie zorientowane na cel wykorzystuje kombinatoryczne regularności warstwy koncepcji do opisywania regularności w strukturze zdarzeń istotnych dla celu. Poszukiwanie przekonania, którego konsekwencja pasuje do opisu celu, jest organizowane przy użyciu struktury kategorii warstwy koncepcji; koncepcje pasują do koncepcji, a nie nieprzeanalizowane obrazy sensoryczne pasujące do nieprzeanalizowanych obrazów sensorycznych. Przeszukiwanie przekonań jest obliczeniowo wykonalne ze względu na wyuczone rezonanse i wyuczone skojarzenia, które same w sobie są „wyuczoną złożonością", a ponadto reprezentują regularności w koncepcyjnie opisanym modelu, a nie w surowym obrazie sensorycznym. Myślenie zorientowane na cel, tak jak jest stosowane przez ludzi, jest często abstrakcyjne, co wymaga wsparcia ze strony właściwości obrazów mentalnych; wymaga, aby umysł utrzymywał opisowe obrazy, które nie są w pełni wizualizowane lub całkowicie zaspokojone przez odniesienie sensoryczne, ale które wiążą się ze specyficznymi odniesieniami, gdy stają się dostępne. Modalności sensoryczne zapewniają przestrzeń, w której może istnieć cała ta wyobraźnia i interpretują środowisko, z którego wyuczona złożoność jest wyuczona. Struktura cech modalności

sprawia, że uczenie się jest obliczeniowo wykonalne. Bez struktury cech koncepcje są obliczeniowo niewykonalne; bez struktury kategorii myśli są obliczeniowo niewykonalne. Bez modalności nie ma doświadczeń ani wyobrażeń mentalnych; bez wyuczonej złożoności nie ma pojęć, które mogłyby ustrukturyzować doświadczenie, ani przekonań uogólnionych z doświadczenia. Oprócz wspierania podstawowych wymagań, modalności przyczyniają się bezpośrednio do inteligencji w każdym przypadku, w którym zachowania referencyjne pokrywają się z zachowaniami modalności, a pośrednio w przypadkach, w których istnieją ważne metafory między zachowaniami modalności i zachowaniami referencyjnymi. Nawet jeśli wymyślenie nowego podcelu jest „wykrzyknikiem myśli" z perspektywy wielu tradycyjnych teorii, jest to wykrzyknik na końcu bardzo długiego zdania. Powstanie pojedynczej myśli jest wydarzeniem, które ma miejsce w całym umyśle – nienaruszonym procesie rozumowania z historią w przeszłości.

Przekonania

Przekonania – wiedza deklaratywna – rozciągają się na granicy między poziomem koncepcji a poziomem myśli. W odniesieniu do cech poziomów wymienionych wcześniej, przekonania są wyuczone, specyficzne, przechowywane i powtarzające się. Z tej perspektywy przekonania należy klasyfikować jako wyuczoną złożoność, a zatem część uogólnionego poziomu koncepcji. Jednak przekonania wykazują większe podobieństwo powierzchniowe do zdań mentalnych niż do pojedynczych słów. Ich wewnętrzna struktura wydaje się przypominać struktury koncepcji bardziej niż koncepcje; a przekonania posiadają cechy, takie jak ustrukturyzowane poprzedniki i następstwa, które są trudne do opisania, z wyjątkiem kontekstu organizacji poziomu myśli. Dlatego zdecydowałem się omówić przekonania na poziomie myśli. Przekonania są nabywane za pośrednictwem dwóch głównych źródeł, indukcji i dedukcji, odnoszących się odpowiednio do uogólnienia na podstawie doświadczenia i rozumowania z poprzednich przekonań. Najsilniejsze przekonania mają zarówno wsparcie indukcyjne, jak i dedukcyjne: wnioski dedukcyjne z potwierdzeniem empirycznym lub indukcyjne uogólnienia z wyjaśnieniami przyczynowymi. Indukcja i dedukcja mogą się przecinać, ponieważ oba obejmują abstrakcję. Generalizacja indukcyjna tworzy opis zawierający kategorie, które działają jak zmienne – abstrakcyjne obrazy, które zmieniają się w zależności od bazy doświadczalnej i ją opisują. Abstrakcyjna dedukcja przyjmuje kilka generalizacji nabytych indukcyjnie lub dedukcyjnie i łączy ze sobą ich abstrakcyjne poprzedniki i abstrakcyjne następstwa, aby wytworzyć abstrakcyjny wniosek, jak zilustrowano we wcześniejszej dyskusji na temat abstrakcyjnych obrazów mentalnych. Nawet całkowicie specyficzne przekonania potwierdzone przez pojedyncze doświadczenie, takie jak „Sylwester Y2K miał miejsce w piątkową noc", są nadal „abstrakcyjne" w tym sensie, że mają oparty na koncepcjach opis struktury kategorii istniejący ponad bezpośrednią pamięcią sensoryczną, a ten opis koncepcyjny można łatwiej powiązać z abstrakcyjnymi przekonaniami, które odnoszą się do tych samych pojęć. Przekonania mogą być sugerowane przez generalizację w oparciu o bazę doświadczalną i wspierane przez generalizację w oparciu o bazę doświadczalną, ale istnieją granice tego, ile wsparcia może wygenerować czysta indukcja (częsta skarga filozofów); zawsze może istnieć przypadek, który je obala, o którym nie wiesz. Generalizacja indukcyjna prawdopodobnie przypomina generalizację pojęć, mniej więcej; istnieje proces początkowego zauważania regularności w całej bazie doświadczalnej, proces jej weryfikacji, a być może nawet proces wytwarzania czegoś podobnego do jąder pojęć w celu wskazywania często istotnych przekonań. Przekonania mają inną strukturę niż koncepcje; koncepcje są albo przydatne, albo nieprzydatne, ale przekonania są albo prawdziwe, albo fałszywe. Koncepcje odnoszą się do odniesień, podczas gdy przekonania opisują relacje między poprzednikami i następstwami. Chociaż oznacza to inny repertuar uogólnień, które wytwarzają indukcyjne przekonania, i inną procedurę weryfikacji, obliczeniowe zadanie zauważania uogólnienia między poprzednikami i następstwami wydaje się silnie przypominać uogólnianie dwumiejscowego predykatu. Przekonania są dobrze znane w tradycyjnej sztucznej inteligencji i często są niebezpiecznie nadużywane; podczas gdy

każdy proces można opisać za pomocą przekonań, nie oznacza to, że proces poznawczy jest implementowany przez przekonania. Posiadam modalność wizualną, która implementuje wykrywanie krawędzi i posiadam przekonania na temat mojej modalności wizualnej, ale ten ostatni aspekt umysłu nie wpływa na pierwszy. Mogę nie posiadać żadnych przekonań na temat wykrywania krawędzi lub mieć zupełnie błędne przekonania na temat wykrywania krawędzi, a moja modalność wizualna będzie nadal działać bez żadnych zakłóceń. AI może być w stanie introspekcji na niższych poziomach organizacji (patrz sekcja 3), a podsystemy poznawcze AI mogą oddziaływać z przekonaniami AI bardziej niż odpowiednie podsystemy u ludzi , ale przekonania i oprogramowanie mózgowe pozostają odrębne – nie tylko odrębne, ale zajmują różne poziomy organizacji. Kiedy szukamy funkcjonalnych konsekwencji przekonań – ich materialnych skutków dla inteligencji AI – powinniśmy szukać wpływu na rozumowanie AI oraz jego późniejsze decyzje i działania. Wszystko można opisać przekonaniem, w tym każde zdarzenie, które ma miejsce w umyśle, ale nie wszystkie zdarzenia w umyśle są implementowane przez posiadanie przekonania, które opisuje reguły rządzące tym zdarzeniem. Kiedy umysł „naprawdę" posiada przekonanie „o" czymś, a nie tylko jakieś niejasne dane, jest częstym pytaniem w filozofii sztucznej inteligencji. Mam coś do powiedzenia na ten temat w następnej sekcji. W formalnych, klasycznych terminach, poznawczy efekt posiadania przekonania jest czasami definiowany w ten sposób, że gdy antecedent przekonania jest spełniony, jego konsekwencja zostaje zakończona. Uważałbym to za jeden z wielu wnioskowań, ale mimo to jest to dobry przykład wnioskowania – poszukiwanie przekonań, których antecedenty są spełnione przez bieżące wyobrażenia, i wnioskowanie z konsekwencji (z poleganiem na samym przekonaniu i na wyobrażeniach dopasowanych do antecedenta). Jednakże owo wnioskowanie, jeśli zostanie zastosowane w ślepym sensie przywoływanym przez logikę klasyczną, doprowadzi do mnóstwa bezużytecznych wniosków; sekwencję należy rozpatrywać w kontekście weryfikatorów, takich jak „Jak rzadko zdarza się, aby to przekonanie było stosowane?", „Jak często to przekonanie jest przydatne, gdy jest stosowane?" lub „Czy wytworzony następnik przecina się z jakimkolwiek innym obrazowaniem, takim jak obrazowanie pytania otwartego?" Niektóre inne sekwencje obejmujące przekonania: Kojarzenie wstecz od obrazowania pytania w celu znalezienia przekonania, którego następnik dotyka obrazowania pytania, a następnie sprawdzenie, czy poprzednik przekonania może zostać spełniony przez bieżące obrazowanie lub ewentualnie przekształcenie poprzednika przekonania w obrazowanie pytania. Znalezienie przekonania przyczynowego, którego następnik odpowiada celowi; poprzednik może stać się podcelem. Wykrycie przypadku, w którym przekonanie jest naruszone – zwykle będzie to bardzo widoczne. Załóżmy, że sztuczna inteligencja z modalnością bilardową indukcyjnie utworzyła przekonanie „wszystkie bilard, które są „czerwone", są „gigantyczne"." Załóżmy dalej, że „czerwony" i „gigantyczny" to koncepcje utworzone przez grupowanie pojedynczych cech, tak że zakres wielkości klastra wskazuje na „gigantyczny", a objętość przestrzeni kolorów klastra wskazuje na „czerwony". Jeśli to przekonanie jest wystarczająco wyraziste, w odniesieniu do bieżącego zadania, aby rutynowo sprawdzać je pod kątem wszystkich obrazów mentalnych, to kilka właściwości poznawczych powinno obowiązywać, jeśli AI naprawdę posiada przekonanie o rozmiarze czerwonych bilardów. W obrazowaniu subjunktywnym, używanym do wyobrażania sobie bilardów niesensorycznych, każdy bilard wyobrażany jako czerwony (w obrębie skupionej objętości kolorów koncepcji „czerwonego") musiałby zostać wyobrażony jako gigantyczny (w obrębie skupionego zakresu rozmiarów koncepcji „gigantycznego"). Jeśli przekonanie „wszystkie czerwone bilardy są gigantyczne" ma wyraźną niepewność, to wniosek o gigantyzmie opierałby się na tym źródle niepewności i dzieliłby postrzeganą wątpliwość. Biorąc pod uwagę zewnętrzne obrazy sensoryczne, jeśli widzi się bilarda, który jest czerwony i mały, należy to postrzegać jako naruszenie przekonania. Biorąc pod uwagę obrazy sensoryczne, jeśli bilard jest w jakiś sposób postrzegany jako „czerwony" przed postrzeganiem jego rozmiaru (trudno sobie wyobrazić, jak to się dzieje u człowieka), to przekonanie musi tworzyć przewidywanie lub oczekiwanie, że bilard będzie gigantyczny, wiążąc wiszącą abstrakcyjną koncepcję

„gigantycznego” z obrazem sensorycznym czerwonego bilarda. Jeśli obraz sensoryczny zostanie ukończony później, a jądro koncepcji „gigantycznego” nie zostanie spełnione przez ukończony obraz sensoryczny czerwonego bilarda, to wynik powinien być naruszonym oczekiwaniem, a ten konflikt powinien rozprzestrzeniać się z powrotem do źródła oczekiwania, aby być postrzeganym jako naruszone przekonanie. Ogólnie rzecz biorąc, przekonania używane w obrazach subjunktywnych kontrolują obrazy bezpośrednio, podczas gdy przekonania używane do interpretowania informacji sensorycznych rządzą oczekiwaniami i określają, kiedy oczekiwanie zostało naruszone. Jednak „sensoryczny” i „subjunktywny” są względne; subjunktywne obrazy rządzone przez jedno przekonanie mogą przecinać się i naruszać inne przekonanie – każde wyobrażenie jest „sensoryczne” w stosunku do przekonania, jeśli to wyobrażenie nie jest bezpośrednio kontrolowane przez to przekonanie. Tak więc abstrakcyjne rozumowanie może wykryć niespójności w przekonaniach. (Niespójność nie powinna powodować, że prawdziwy umysł wrzeszczy z przerażenia i się załamuje, ale powinna być znaczącym wydarzeniem, które zmienia tok myślenia na poszukiwanie źródła niespójności, przyglądając się przekonaniom i twierdzeniom, na których się opiera, i sprawdzając ich pewność. Wykrycia niespójności, wyrażone jako myśli, mają tendencję do tworzenia obrazów pytań i celów wiedzy, które kierują rozważania w stronę rozwiązania niespójności.) Współewolucja myśli i języka: Początki wewnętrznej

Narracja

Dlaczego transformacja struktur pojęciowych w liniowe sekwencje słów, oczywiście niezbędna do komunikacji mówionej, odbywa się również w wewnętrznym strumieniu świadomości? Dlaczego nie używać tylko struktur pojęciowych? Dlaczego transformujemy struktury pojęciowe w zdania gramatyczne, jeśli nikt nie słucha? Czy jest to konieczna część inteligencji? Czy sztuczna inteligencja musi robić to samo, aby funkcjonować? Spór o to, co było pierwsze, myśl czy język, jest starożytny w filozofii. Współcześni badacze ewolucji języka próbują rozbić ewolucję języka na stopniowo adaptacyjne etapy, opisać wiele funkcji, które są łącznie wymagane dla języka i wyjaśnić, w jaki sposób mogły powstać preadaptacje dla tych funkcji. Dekompozycje funkcjonalne unikają niektórych paradoksów typu „co było pierwsze: kura czy jajko”, które wynikają z postrzegania języka jako funkcji monolitycznej. Niestety, istnieją dalsze paradoksy wynikające z postrzegania języka niezależnie od myśli lub z postrzegania myśli jako funkcji monolitycznej. Z perspektywy teoretyka kognitywnego język jest tylko jedną z funkcji współczesnego ludzkiego supersystemu poznawczego, ale z perspektywy teoretyka ewolucyjnego cechy językowe determinują, które naciski selekcji społecznej dotyczą ewolucji poznania w danym punkcie. Stąd „współewolucja myśli i języka”, a nie „ewolucja języka jako jednej części myśli”. Ewolucyjne wyjaśnienie samego języka zostanie „zablokowane”, gdy po raz pierwszy osiągnie cechę, która jest adaptacyjna dla poznania i preadaptacyjna dla języka, ale dla której nie istnieje niezależna presja selekcji językowej w przypadku braku już istniejącego języka. Ponieważ obecnie nie ma konsensusu co do funkcjonalnego rozkładu inteligencji, współcześni teoretycy ewolucji języka czasami nie są w stanie uniknąć takich punktów spornych. Na pierwszy rzut oka DGI może wydawać się wyjaśniać ewolucyjność języka jedynie poprzez rozróżnienie poziomu pojęć i poziomu myśli; dopóki istnieją proste odruchy wykorzystujące wyuczoną strukturę kategorii, opracowanie poziomu konceptualnego będzie niezależnie adaptacyjne, nawet w przypadku braku ludzkiego poziomu myśli. Opracowanie poziomu konceptualnego w celu wsparcia skojarzeń międzymodalnych wydaje się umożliwiać przekraczanie luki między sygnałem a koncepcją, a opracowanie poziomu konceptualnego w celu wsparcia mieszania lub łączenia pojęć (adaptacyjne, ponieważ umożliwia organizmowi postrzeganie prostych regularności kombinatorycznych) wydaje się umożliwiać prymitywne, niesyntaktyczne sekwencje słów. Ogólnie rzecz biorąc, przypomina to obraz Bickertona protojęzyka jako ewolucyjnego pośrednika, w którym wyuczone sygnały przekazują wyuczone pojęcia, a wiele pojęć miesza się, ale bez składni przekazującej informacje docelowe. Gdy istniał protojęzyk,

mogły przejąć kontrolę właściwe naciski selekcji językowej. Jednak, jak wskazuje , obraz ten nie wyjaśnia, dlaczego inne gatunki nie rozwinęły protojęzyka. Powiązanie międzymodalne nie ogranicza się do ludzi ani nawet naczelnych. Deacon sugeruje, że niektóre niezbędne kroki myślowe w języku są nie tylko nieintuicyjne, ale wręcz kontrintuicyjne dla gatunków nieludzkich, tak samo jak test selekcji Wasona jest kontrintuicyjny dla ludzi. Opis tego „niezręcznego kroku" autorstwa Deacona wykorzystuje inną teorię inteligencji jako tło, a ja miałbym zatem inne spojrzenie na naturę niezręcznego kroku: podejrzewam, że szympansy mają niezwykłe trudności z nauką symboli, tak jak je rozumiemy, ponieważ język, nawet protojęzyk, wymaga tworzenia abstrakcyjnych obrazów mentalnych, które mogą wisieć bez wsparcia, a następnie wiązać się z później napotkanym odniesieniem sensorycznym. Kluczowa trudność w języku – krok, który jest niezręczny dla innych gatunków – nie polega na zdolności kojarzenia sygnałów; naczelne (i szczury, jeśli o to chodzi) mogą łatwo skojarzyć sygnał percepcyjny z wymaganym działaniem lub stanem świata. Niezręczny krok polega na tym, że sygnał wywołuje kategorię jako abstrakcyjne obrazy, niezależnie od bezpośrednich odniesień sensorycznych, które mogą wiązać się z później napotkanym odniesieniem. Ten krok jest dla nas całkowicie rutynowy, ale mógłby być niemal niemożliwy w przypadku braku wsparcia projektowego dla „zawieszania koncepcji w powietrzu". W przypadku braku myśli istnieje niewiele powodów, dla których gatunek uznałby za przydatne zawieszanie koncepcji w powietrzu. W przypadku braku języka istnieje jeszcze mniej powodów, aby skojarzyć sygnał percepcyjny z przywołaniem koncepcji jako abstrakcyjnego obrazu. Język jest trudny dla innych gatunków, nie z powodu luki między sygnałem a koncepcją, ale dlatego, że język wykorzystuje cechę obrazowania mentalnego, dla której nie ma wystarczającego wsparcia projektowego u innych gatunków. Podejrzewam, że mógł to być kontekst adaptacyjny dla abstrakcyjnych obrazów, a nie naciski selekcji językowej, co doprowadziło do adaptacji, która okazała się preadaptacyjna dla symbolizacji i stąd zaczęła niektóre gatunki naczelnych zsuwać się w dół gradientu dostosowania, który obejmował koewolucję myśli i języka. Jeśli, jak sugeruje ten obrazek, ewolucja przedczłowieczakowata przede wszystkim rozwinęła warstwę pojęciową (w sensie rozwijania procesów mózgowych, które wspierają kategorie, a nie w sensie dodawania wyuczonych pojęć jako takich), to oznacza, że warstwa pojęciowa może zawierać większość wspierającej złożoności funkcjonalnej dla ludzkiego poznania. Nie wynika to koniecznie, ponieważ ewolucja mogła spędzić dużo czasu, ale niewiele uzyskać w zamian, ale jest to przynajmniej sugestywne. (Ta sekcja na poziomie pojęć jest w rzeczywistości najdłuższą sekcją.) Powyższy obrazek sugeruje również, że rodzina człekokształtnych mogła współewoluować kombinatoryczne struktury pojęciowe, które wewnętrznie modyfikują obrazy mentalne (myśli) i kombinatoryczne struktury pojęciowe, które wywołują obrazy mentalne u osobników tego samego gatunku (język). Oczywiste jest, że język wykorzystuje wiele funkcji pierwotnie opracowanych w celu wspierania wewnętrznego poznania, ale koewolucja myśli i języka oznacza odpowiednią okazję do ewolucyjnego opracowania myśli hominidów w celu zawłaszczenia funkcji pierwotnie rozwiniętych w celu wspierania języka hominidów. Pozorna konieczność wewnętrznej narracji dla ludzkich rozważań może okazać się introspekcyjną iluzją, ale jeśli jest prawdziwa, to silnie sugeruje, że funkcjonalność językowa została zawłaszczona dla funkcjonalności poznawczej podczas ewolucji człowieka. Cechy językowe, takie jak specjalne przetwarzanie tagów, które wywołują koncepcje, lub stosowanie składni w celu organizowania złożonych wewnętrznych informacji docelowych dla struktur pojęć kombinatorycznych, mogą być również adaptacyjne lub preadaptacyjne dla efektywnego myślenia. Tylko kilka takich cech językowych musiałoby zostać zawłaszczonych jako niezbędne części myśli, zanim „strumień świadomości" stałby się utrwaloną częścią ludzkiej inteligencji. Jest to prawdopodobnie wystarczające wyjaśnienie istnienia wewnętrznej narracji, prawdopodobnie czyniąc wewnętrzną narrację czystym spandrelem (cechą wyłaniającą się, ale nieadaptacyjną). Jednak ostrożność w AI, a nie ostrożność w psychologii ewolucyjnej, powinna skłonić nas do zastanowienia się, czy nasza wewnętrzna narracja pełni funkcję adaptacyjną. Na przykład nasza wewnętrzna narracja mogłaby wyrażać rozważania w

formie, którą możemy łatwiej przetworzyć jako (wewnętrzne) doświadczenie sensoryczne na potrzeby introspekcji i pamięci; lub proces poznawczy narzucania wewnętrznych myśli na obrazy mentalne mógłby zawłaszczyć mechanizm językowy, który również tłumaczy komunikację zewnętrzną na obrazy mentalne; lub wewnętrzna narracja może zawłaszczyć inteligencję społeczną, która modeluje innych ludzi poprzez odnoszenie się do ich komunikacji, w celu modelowania siebie. Ale nawet jeśli ewolucja hominidów zawłaszczyła narrację wewnętrzną, ogólny model nadal sugeruje, że – chociaż nie możemy oddzielić języka od inteligencji ani oddzielić ewolucji myśli od ewolucji języka – de novo projekt umysłu mógłby oddzielić inteligencję od języka. To z kolei sugeruje, że SI mogłaby używać struktur pojęciowych bez serializowania ich jako zdań gramatycznych tworzących wewnętrzną narrację języka naturalnego, o ile wszystkie funkcjonalności językowe zapożyczone dla ludzkiej inteligencji byłyby odtwarzane w terminach niejęzykowych – w tym wyrażanie myśli w formie dostępnej introspektywnie i stosowanie złożonego wewnętrznego targetowania w strukturach pojęciowych. Obserwowanie SI może wymagać rejestrowania myśli SI i tłumaczenia tych myśli na formy zrozumiałe dla człowieka, a programiści mogą musieć przekazywać SI struktury pojęciowe, ale nie musi to oznaczać, że SI jest zdolna do rozumienia lub tworzenia języka ludzkiego. Prawdziwa komunikacja językowa między ludźmi a SI może pojawić się znacznie później w rozwoju, być może jako zwykła kompetencja domenowa, a nie talent wspierany przez oprogramowanie mózgowe. Oczywiście, rozumienie języka ludzkiego i naturalna ludzka konwersacja to niezwykle atrakcyjny cel i niewątpliwie zostałyby podjęte tak wcześnie, jak to możliwe; jednak wydaje się, że język nie musi być wdrażany natychmiast lub jako niezbędny warunek wstępny rozważań.

Poziom rozważań

Od myśli do rozważań

U ludzi wyższe poziomy organizacji są na ogół bardziej dostępne dla introspekcji. Nie jest zaskakujące, że wewnętrzne zdarzenia poznawcze zwane „myślami", jak opisano w poprzedniej sekcji, wydają się dziwnie znajome; słuchamy myśli przez cały dzień. Niebezpieczeństwo dla programistów AI polega na tym, że treść poznawcza, która jest otwarta na introspekcję, jest czasami kusząco łatwa do bezpośredniego przetłumaczenia na kod. Ale jeśli ludzie rozwinęli cykliczną interakcję myśli i obrazów, sam ten fakt nie dowodzi (ani nawet nie argumentuje), że projekt jest dobry. Jaka jest materialna korzyść dla inteligencji z używania tablicowych obrazów umysłowych i sekwencji, zamiast prostszych stałych algorytmów „rozumowania" w klasycznej AI? Ewolucja charakteryzuje się rosnącymi poziomami organizacji o coraz większym rozwinięciu, złożoności, elastyczności, bogactwie i kosztach obliczeniowych; złożoność wyższych warstw nie wyłania się automatycznie wyłącznie z dolnej warstwy, ale podlega presji selekcyjnej i ewolucji złożonej adaptacji funkcjonalnej – adaptacji, która jest istotna na tym poziomie i, jak się okazuje, czasami preadaptacyjna dla wyłaniania się wyższych poziomów organizacji. Ta sygnatura projektowa wyłania się przynajmniej częściowo z charakterystycznej ślepoty ewolucji i może nie być koniecznym idiomem umysłów w ogóle. Niemniej jednak poprzednie próby bezpośredniego programowania zjawisk poznawczych, które powstają na postmodalnych poziomach organizacji, zakończyły się całkowitą porażką. Istnieją specyficzne patologie AI, które wyłaniają się z tej próby, takie jak problem uziemienia symboli i problem zdrowego rozsądku. U ludzi koncepcje są płynnie elastyczne i ekspresyjne, ponieważ powstają z modalności; myśli są płynnie elastyczne i ekspresyjne, ponieważ powstają z koncepcji. Nawet biorąc pod uwagę wartość obrazów tablicy i sekwencji w izolacji – na przykład, biorąc pod uwagę architekturę AI, która używała stałych algorytmów deliberacji, ale używała tych algorytmów do tworzenia i wywoływania myśli DGI – nadal istnieją konieczne powody, dla których wzorce deliberatywne muszą być budowane na zachowaniach poziomu myśli, a nie implementowane jako niezależny kod; istnieją patologie AI, które wynikałyby z próby implementacji deliberacji w sposób czysto odgórny. Istnieje złożoność odgórna w deliberacji –

adaptacyjna funkcjonalność, którą najlepiej postrzegać jako odnoszącą się do poziomu deliberacji, a nie poziomu myśli – ale ta złożoność jest w większości ucieleśniona jako zachowania poziomu myśli, które wspierają wzorce deliberatywne. Ponieważ poziom deliberacji elastycznie wyłania się z sekwencji poziomu myśli, ciąg myśli można przekierować bez ich zniszczenia. Aby użyć przykładu podanego wcześniej, jeśli umysł deliberatywny zastanawia się „Dlaczego X jest Y?", ale nie znajduje wyjaśnienia, ta lokalna porażka nie jest katastrofą dla deliberacji jako całości. Umysł może mentalnie odnotować pytanie jako nierozwiązaną zagadkę i kontynuować z innymi sekwencjami. Naruszenie przekonania nie niszczy umysłu; staje się ono przedmiotem uwagi i kolejną rzeczą do rozważenia. Odkrycie niespójnych przekonań nie powoduje załamania, jak miałoby to miejsce w systemie logiki monotonicznej, ale zamiast tego przesuwa punkt ciężkości uwagi na sprawdzanie i rewidowanie logiki dedukcyjnej. Rozważanie splata wiele przecinających się wątków rozumowania poprzez przecinające się obrazy, przy czym przystanki, a nawet cel końcowy nie zawsze są znane z góry. We wszechświecie złych programów telewizyjnych wypowiedzenie paradoksu Epimenidesa „To zdanie jest fałszywe" do sztucznego umysłu powoduje, że umysł ten krzyczy z przerażenia i zapada się w stertę tlących się części. Opiera się to na stereotypie procesów myślowych, które nie mogą odwrócić, nie mogą zatrzymać i nie posiadają żadnej oddolnej zdolności do zauważania regularności w rozszerzonej sekwencji myśli. Biorąc pod uwagę, w jaki sposób rozważania wyłaniają się z poziomu myśli, można sobie wyobrazić wystarczająco wyrafinowaną, wystarczająco refleksyjną SI, która mogłaby naturalnie przezwyciężyć paradoks Epimenidesa. Napotkanie paradoksu „To zdanie jest fałszywe" prawdopodobnie rzeczywiście doprowadziłoby początkowo do zapętlonej sekwencji myśli, ale nie spowodowałoby to trwałego zablokowania SI; zamiast tego doprowadziłoby do kategoryzacji w powtarzających się myślach (jak człowiek zauważający paradoks po kilku cyklach), która kategoryzacja stałaby się wówczas widoczna i mogłaby być rozważana sama w sobie przez inne sekwencje. Jeśli SI jest wystarczająco kompetentna w rozumowaniu dedukcyjnym i introspektywnej generalizacji, może uogólniać konkretne przypadki „Jeśli stwierdzenie jest prawdziwe, musi być fałszywe" i „Jeśli stwierdzenie jest fałszywe, musi być prawdziwe" jako dwie ogólne klasy myśli wytworzone przez paradoks i pokazać, że rozumowanie z myśli jednej klasy prowadzi do myśli innej klasy; jeśli tak, SI może wywnioskować – nie tylko indukcyjnie zauważyć, ale dedukcyjnie współtwierdzić – że proces myślowy jest wieczną pętlą. Oczywiście nie dowiemy się, czy to naprawdę działa w ten sposób, dopóki tego nie spróbujemy. Użycie modelu sequitur tablicy nie jest automatycznie wystarczające do głębokiej refleksyjności; SI, która posiadałaby ograniczony repertuar sequitur, brak refleksyjności, brak zdolności do stosowania refleksyjnej kategoryzacji i brak zdolności do zauważania, kiedy ciąg myśli nie przyniósł niczego użytecznego przez jakiś czas, mogłaby nadal wiecznie pętlić się przez paradoks jako wyłaniający się, ale bezużyteczny produkt repertuaru sequitur. Przekroczenie paradoksu Epimenidesa wymaga umiejętności przeprowadzania indukcyjnej generalizacji i dedukcyjnego rozumowania na podstawie doświadczeń introspekcyjnych. Wymaga to jednak również organizacji oddolnej w rozważaniach, tak aby spontaniczna introspekcyjna generalizacja mogła przyciągnąć uwagę. Rozważania muszą wynikać z myśli, a nie tylko wykorzystywać myśli do wdrażania sztywnych algorytmów. Osiągnąwszy poziom rozważań, w końcu odwracamy się od długiego opisu tego, czym jest umysł, i skupiamy się na tym, co umysł robi – użytecznych operacjach wdrażanych przez sekwencje myśli, które są strukturami pojęć, które są abstrahowane od doświadczenia sensorycznego w modalnościach sensorycznych.

Wymiary inteligencji

Filozofowie często definiują „prawdę" jako porozumienie między wiarą a rzeczywistością; formalnie jest to znane jako „teoria korespondencji" prawdy. W ramach teorii korespondencji prawdy filozofowie sztucznej inteligencji często definiowali „wiedzę" jako mapowanie między wewnętrznymi strukturami danych a zewnętrzną rzeczywistością fizyczną . Rozpatrywana w izolacji teoria korespondencji wiedzy jest łatwo nadużywana; można jej używać do argumentowania na podstawie mapowań, które okazują

się istnieć wyłącznie w umyśle programisty. Inteligencja jest ewolucyjną zaletą, ponieważ umożliwia nam modelowanie, przewidywanie i manipulowanie rzeczywistością. Mówiąc to, nie opowiadam się za stanowiskiem filozoficznym, że tylko użyteczna wiedza może być prawdziwa. Istnieje wystarczająca regularność w aktywności zdobywania wiedzy, w szerokim spektrum problemów wymagających wiedzy, że ewolucja ma tendencję do tworzenia niezależnych sił poznawczych do poszukiwania prawdy. Najlepiej postrzegać poszczególne organizmy jako wykonawców adaptacji, a nie maksymalizatorów sprawności. „Poszukiwanie prawdy", nawet gdy jest postrzegane jako zaledwie lokalne podzadanie większego problemu, ma wystarczającą autonomię funkcjonalną, że wiele adaptacji ludzkich jest lepiej postrzeganych jako „poszukiwanie prawdy" niż „poszukiwanie użytecznych przekonań". Ponadto, zgodnie z moją własną filozofią, powiedziałbym, że przekonania są użyteczne, ponieważ są prawdziwe, a nie „prawdziwe", ponieważ są użyteczne. Ale użyteczność jest silniejszym i bardziej wiarygodnym testem prawdy; trudniej ją oszukać. Społeczny proces nauki stosuje przewidywanie jako test modeli, a te same modele, które dają udane przewidywania, są często wystarczająco dobrymi przybliżeniami do konstruowania technologii (manipulacji). Rozróżniłbym cztery kolejne, silniejsze stopnie wiązania między modelem a rzeczywistością:

• Wiązanie sensoryczne występuje, gdy istnieje mapowanie między treścią poznawczą w modelu a cechami rzeczywistości zewnętrznej. Bez testów użyteczności nie ma formalnego sposobu na zapobieganie nadużyciom domniemanych wiązań sensorycznych; domniemane mapowanie może leżeć głównie w umyśle obserwatora. Jednak jeśli system jako całość przechodzi testy użyteczności, duża część zadania rozszerzania i ulepszania modelu nadal będzie lokalnie polegać na odkrywaniu dobrych wiązań sensorycznych – znajdowaniu przekonań, które są prawdziwe zgodnie z intuicyjną „teorią korespondencji" prawdy.

• Wiązanie predykcyjne występuje, gdy model może być użyty do prawidłowego przewidywania przyszłych zdarzeń. Z wewnętrznej perspektywy AI, wiązanie predykcyjne występuje, gdy model może być użyty do prawidłowego przewidywania przyszłych danych sensorycznych. AI może zostać wezwana do dokonywania udanych przewidywań dotyczących rzeczywistości zewnętrznej (poza komputerem), wirtualnych mikrośrodowisk (wewnątrz komputera, ale poza AI) lub wyniku procesów poznawczych (wewnątrz AI, ale przebiegających oddzielnie od przewidywania). „Informacje sensoryczne" mogą pochodzić nie tylko z urządzenia sensorycznego ukierunkowanego na rzeczywistość zewnętrzną, ale także z poznania sensorycznego ukierunkowanego na dowolny proces, którego wynik na przewidywanym poziomie nie podlega bezpośredniej kontroli. (Oczywiście, z naszej perspektywy, przewidywanie „rzeczywistego świata" pozostaje najsilniejszym testem.)

• Decydujące wiązanie występuje, gdy model może przewidzieć skutki kilku możliwych działań na rzeczywistość i wybrać działanie, które daje najlepszy wynik w ramach pewnego systemu celów (patrz poniżej). Przewidując wyniki w ramach kilku możliwych stanów świata, składających się z obecnego stanu świata plus każdego z kilku możliwych działań, możliwe staje się wybieranie między przyszłościami.

• Manipulacyjne wiązanie występuje, gdy AI może opisać pożądaną przyszłość za pomocą obrazów subjunktywnych i wymyślić sekwencję działań, która prowadzi do tej przyszłości. Podczas gdy decyzja obejmuje wybranie jednego działania z ustalonego i ograniczonego zestawu, manipulacja obejmuje wymyślanie nowych działań, być może działań wcześniej niezauważanych, ponieważ zestaw możliwych działań jest nieograniczony lub obliczeniowo duży. Najprostszą formą manipulacji jest odwoływanie się wstecz do celów rodzica i dziecka, wykorzystując w tym celu przekonania przyczynowe; nie jest to jedyna forma manipulacji, ale jest lepsza niż wyczerpujące przeszukiwanie wszystkich możliwych działań.

Wyróżniam także trzy kolejne stopnie zmiennej złożoności:

• Zmienna dyskretna ma referenty wybrane z ograniczonego zbioru, który jest obliczeniowo mały – na przykład zbiór 20 możliwych akcji lub zbiór 26 możliwych małych liter. Binarna obecność lub brak cechy jest również zmienną dyskretną.

• Zmienna ilościowa jest wybierana z zbioru liczb rzeczywistych lub z obliczeniowo dużego zbioru, który przybliża płynnie zmieniającą się wielkość skalarną (taką jak zbiór liczb zmiennoprzecinkowych).

• Zmienna wzorzysta składa się ze skończonej liczby elementów ilościowych lub dyskretnych. Przykłady: Skończony ciąg małych liter, np. „mkrznye". Rzeczywisty punkt w przestrzeni 3D (trzy elementy ilościowe). Dwuwymiarowy czarno-biały obraz (dwuwymiarowa tablica pikseli binarnych).

Wymiar złożoności zmiennej jest ortogonalny do wymiaru SPDM (sensorypredictive- decision-manipulative), ale podobnie jak SPDM opisuje on kolejno trudniejsze testy inteligencji. Decydujące powiązanie pożądanego wyniku z pożądanym działaniem jest wykonalne obliczeniowo tylko wtedy, gdy „działanie" jest zmienną dyskretną wybraną z małego zestawu – wystarczająco małego, aby można było modelować każde możliwe działanie. Gdy działanie jest zmienną ilościową, wybraną z dużych obliczeniowo zestawów, takich jak liczby zmiennoprzecinkowe w przedziale [0, 1], konieczna jest jakaś forma manipulacyjnego powiązania, taka jak łańcuchowanie wsteczne, aby dojść do określonego wymaganego działania. (Należy zauważyć, że dodanie ciągłego parametru czasu do dyskretnego działania czyni je ilościowym). Powiązanie precyzyjnego ilościowego obrazu celu z precyzyjnym działaniem ilościowym nie może zostać wykonane poprzez wyczerpujące testowanie alternatyw; wymaga sposobu na przekształcenie obrazu celu tak, aby dojść do obrazu podcelu lub obrazu działania. Najprostszą transformacją jest relacja tożsamości – ale nawet transformacja tożsamości nie jest możliwa w przypadku mechanizmu czysto do przodu. Kolejną najprostszą metodą byłoby zastosowanie przekonania przyczynowego, które określa odwracalną relację między poprzednikiem a następstwem. W zadaniach sterowania w czasie rzeczywistym modalności ruchowe (u ludzi cały układ sensomotoryczny) mogą automatycznie wytwarzać symfonie działań w celu osiągnięcia celów ilościowych lub wzorcowanych. Ciąg kilku zmiennych dyskretnych lub ilościowych tworzy zmienną wzorcowaną, która prawdopodobnie będzie również obliczeniowo niemożliwa do wykonania w wyczerpującym wyszukiwaniu do przodu. Powiązanie wzorcowego celu ze wzorcowanym działaniem, jeśli relacja nie jest relacją bezpośredniej tożsamości, wymaga (ponownie) przekonania przyczynowego, które określa odwracalną relację między poprzednikiem a następstwem, lub (jeśli takie przekonanie nie nadchodzi) deliberatywnej analizy złożonych prawidłowości w relacji między działaniem a wynikiem, lub eksploracyjnego dostrajania, po którym następuje indukcja, na podstawie której dostrajanie zwiększa pozorne podobieństwo między wynikiem a pożądanym wynikiem. Istnieją poziomy organizacji w ramach powiązań; luźne wiązanie na jednym poziomie może dać początek ściślejszemu wiązaniu na wyższym poziomie. Pręciki i czopki siatkówki odpowiadają przychodzącym fotonom, które odpowiadają punktom na powierzchni obiektu. Wiązanie między metaforycznym pikselem w siatkówce a punktem na powierzchni rzeczywistego świata jest bardzo słabe, bardzo kruche; błądzący promień światła może gwałtownie zmienić wykrytą intensywność optyczną. Ale rzeczywiste doświadczenie sensoryczne zajmuje jeden poziom organizacji powyżej pojedynczych pikseli. Kruche wiązanie sensoryczne między pikselami siatkówki a punktami powierzchni, na niższym poziomie organizacji, daje początek stałemu wiązaniu sensorycznemu między naszym postrzeganiem całego obiektu a samym obiektem. Dopasowanie między dwiema zmiennymi dyskretnymi lub dwiema przybliżonymi zmiennymi ilościowymi może powstać przypadkowo; dopasowanie między dwiema wzorzystymi zmiennymi na wyższym holonicznym poziomie organizacji jest znacznie mniej prawdopodobne, aby powstało z całkowitego zbiegu okoliczności, chociaż może wynikać z przyczyny innej niż oczywista. Jądra koncepcji w ludzkim rozpoznawaniu wzrokowym również wiążą się z całym

doświadczeniem percepcyjnym obiektu, a nie z poszczególnymi pikselami obiektu. Na jeszcze wyższym poziomie organizacji manipulacyjne powiązanie między ludzką inteligencją a światem rzeczywistym jest przybite wieloma indywidualnie ścisłymi powiązaniami sensorycznymi między obrazami koncepcyjnymi a odniesieniami ze świata rzeczywistego. W ludzkiej implementacji istnieją co najmniej trzy poziomy organizacji w ramach teorii korespondencji prawdy! Patologia AI, którą postrzegamy jako „słabą semantykę" – co jest bardzo trudne do zdefiniowania, ale jest intuicyjnym wrażeniem podzielanym przez wielu filozofów AI – może wynikać z pominięcia poziomów organizacji w powiązaniu między modelem a jego odniesieniem.

Akcje

Seria akcji motorycznych, których używam, aby uderzyć w klawisz na mojej klawiaturze, ma wystarczająco dużo stopni swobody, aby „który klawisz naciskam" jako zmienna dyskretna lub „sekwencja uderzonych klawiszy" jako zmienna wzorcowa, podlegały bezpośredniej specyfikacji. Nie muszę angażować się w złożone planowanie, aby uderzyć w sekwencję klawiszy „hello world" lub „labm4"; mogę określić słowa lub litery bezpośrednio i bez potrzeby złożonego planowania. Moje obszary motoryczne i móżdżek wykonują ogromną ilość pracy za kulisami, ale jest to praca, która została zoptymalizowana do punktu subiektywnej niewidoczności. Naciśnięcie klawisza jest zatem działaniem dla celów pragmatycznych, chociaż dla początkującego maszynisty może być celem. Jako pierwsze przybliżenie, obrazowanie celu zostało zredukowane do obrazowania działania, gdy obrazowanie może kierować umiejętnością w czasie rzeczywistym w odpowiedniej modalności. Nie oznacza to koniecznie, że akcje są przekazywane umiejętnościom bez dalszej interakcji; manipulacje w czasie rzeczywistym czasami idą źle, w takim przypadku współzależność między celami a działaniami i umiejętnościami staje się bardziej skomplikowana, czasami z wieloma zmieniającymi się celami wchodzącymi w interakcję z umiejętnościami w czasie rzeczywistym. Obrazowanie zbliża się do poziomu działania, gdy staje się w stanie wchodzić w interakcję z umiejętnościami w czasie rzeczywistym. Czasami cel nie sprowadza się bezpośrednio do działań, ponieważ odniesienie do celu jest fizycznie odległe lub fizycznie oddzielone od „efektorów" – przydatków motorycznych lub ich wirtualnych odpowiedników – tak że manipulowanie odniesieniem do celu zależy najpierw od przezwyciężenia fizycznego oddzielenia jako podproblemu. Jednak w rutynowej aktywności współczesnych ludzi innym bardzo powszechnym powodem, dla którego obrazowanie celu nie przekłada się bezpośrednio na obrazowanie działania, jest to, że obrazowanie celu jest abstrakcyjną cechą wysokiego poziomu, poznawczo oddzieloną od sfery bezpośrednich działań. Mogę kontrolować każde naciśnięcie klawisza podczas pisania, ale ilościowe postrzeganie jakości pisania, do którego odnosi się obrazowanie celu wysokiej jakości pisania, nie podlega bezpośredniej manipulacji. Nie mogę bezpośrednio ustawić jakości mojego pisma na równą Szekspirowi, w sposób, w jaki mogę bezpośrednio ustawić naciśnięcie klawisza na równe „H", ponieważ jakość pisma jest pochodną, abstrakcyjną wielkością. Lepszym słowem niż „abstrakcyjny" jest „holoniczny", termin użyty wcześniej w [51], używany do opisania sposobu, w jaki pojedyncza jakość może być jednocześnie całością złożoną z części i częścią większej całości. Jakość pisma jest ilościowym holonem, który jest ostatecznie związany z serią dyskretnych naciśnięć klawiszy. Mogę bezpośrednio wybierać naciśnięcia klawiszy, ale nie mogę bezpośrednio wybierać holonu jakości pisma. Aby zwiększyć jakość pisma akapitu, muszę połączyć holon jakości pisma z holonami niższego poziomu, takimi jak poprawna pisownia i pomijanie zbędnych słów, które są cechami holonów zdań, które są tworzone poprzez działania naciśnięć klawiszy. Obrazowanie akcji jest zazwyczaj, choć nie zawsze, poziomem, na którym zmienne są całkowicie wolne (bezpośrednio określalne z wieloma stopniami swobody); wyższe poziomy obejmują interakcyjne ograniczenia, które muszą zostać rozwiązane poprzez rozważania.

Cele

Bardzo abstrakcyjne obrazy celów dla jakości pisania są powiązane z bezpośrednio określonymi obrazami działań dla słów i uderzeń w klawisze poprzez pośrednią serię celów podrzędnych, które dziedziczą pożądanie od celów nadrzędnych. Ale czym są cele? Czym jest pożądanie? Do tej pory używałem intuicyjnej definicji tych terminów, która często wystarczała do opisania interakcji systemu celów z innymi systemami, ale nie jest opisem samego systemu celów. Niestety, ludzki system celów jest nieco... zagmatwany... jak wiesz, jeśli jesteś człowiekiem. Większość ludzkiego systemu celów pierwotnie ewoluowała w nieobecności inteligencji deliberatywnej, a w rezultacie zachowania, które przyczyniają się do przetrwania i reprodukcji, mają tendencję do ewoluowania jako niezależne popędy. Przyjmując intencjonalne stanowisko wobec ewolucji, powiedzielibyśmy, że popęd seksualny jest celem reprodukcji dziecka. W czasie ewolucji może to być słuszne stanowisko. Jednak poszczególne organizmy są najlepiej postrzegane jako wykonawcy adaptacji, a nie maksymalizatorzy sprawności, a popęd seksualny nie jest poznawczo celem reprodukcji dziecka; stąd współczesne stosowanie antykoncepcji. Dalsze komplikacje wprowadzane są na poziomie naczelnych przez istnienie złożonych grup społecznych; w konsekwencji naczelne mają „moralne" adaptacje, takie jak wzajemny altruizm, interwencja osób trzecich w celu rozwiązania konfliktów („troska o społeczność") i moralistyczna agresja wobec przestępców społecznych . Jeszcze dalsze komplikacje wprowadzane są przez istnienie rozumowania deliberatywnego i komunikacji językowej u ludzi; ludzie są niedoskonałymi, zwodniczymi organizmami społecznymi, które spierają się o motywy innych w kontekstach adaptacyjnych. Wytworzyło to to, co mogę nazwać jedynie „filozoficznymi" adaptacjami, takimi jak sposoby, w jakie rozumujemy o przyczynowości w argumentach moralnych – ostatecznie dając nam zdolność do wydawania (negatywnego!) osądu na temat wartości moralnej naszych ewoluujących systemów celów i samej ewolucji. Nie zamierzam rozplątywać tej rozległej sieci przyczynowości w tym artykule, chociaż pisałem (nieformalnie, ale obszernie) o tym problemie gdzie indziej, w tym opis architektur poznawczych i motywacyjnych wymaganych, aby umysł angażował się w tak pozornie paradoksalne zachowania, jak wydawanie spójnych osądów na temat własnych celów najwyższego poziomu. (Na przykład umysł może postrzegać bieżącą reprezentację moralności jako probabilistyczne przybliżenie do odniesienia moralnego, o którym można wnioskować). Architektura moralności to dążenie, które idzie w parze z dążeniem do ogólnej inteligencji, a te dwa nie powinny być rozdzielane z powodów, które powinny być oczywiste i staną się jeszcze bardziej oczywiste w sekcji 3; ale niestety po prostu nie ma wystarczająco dużo miejsca, aby zająć się tymi problemami tutaj. Zauważę jednak, że ludzki system celów czasami robi Złą Rzecz31 i nie sądzę, aby AI powinna podążać tymi śladami; umysł może dzielić nasz moralny układ odniesienia, nie będąc funkcjonalnym duplikatem ludzkiego supersystemu celów. W tym artykule odłożę na bok kwestię rozumowania moralnego i przyjmę za pewnik, że system wspiera treść moralną. Pytanie brzmi zatem, w jaki sposób treść moralna wiąże się z obrazami celu i ostatecznie z działaniami. Obrazowanie opisujące supercel jest treścią moralną i opisuje zdarzenia lub stany świata, które umysł uważa za mające wartość wewnętrzną. W terminologii klasycznej opis supercelu jest analogiczny do funkcji użyteczności wewnętrznej. Klasycznie, całkowita użyteczność zdarzenia lub stanu świata to jego użyteczność wewnętrzna, powiększona o sumę użyteczności wewnętrznej (dodatniej lub ujemnej) przyszłych zdarzeń, do których przewiduje się, że to zdarzenie doprowadzi, pomnożoną w każdym przypadku przez przewidywane prawdopodobieństwo przyszłego zdarzenia jako konsekwencji. (Należy zauważyć, że przewidywane konsekwencje obejmują zarówno bezpośrednie, jak i pośrednie konsekwencje, tj. konsekwencje konsekwencji są uwzględniane w sumie). Na pierwszy rzut oka może się to wydawać kolejną zbyt uproszczoną definicją dobrej staromodnej sztucznej inteligencji, ale tym razem będę argumentować na jej korzyść; klasyczna definicja jest bardziej owocna w przypadku złożonych zachowań niż mogłoby się wydawać na pierwszy rzut oka. Pożądanie własności powinno być współrozległe z własnością, która według przewidywań prowadzi do wewnętrznej użyteczności, i zachowywać się identycznie. Określenie, które działania mają prowadzić do największej całkowitej wewnętrznej użyteczności, oraz wymyślanie działań, które

prowadzą do większej wewnętrznej użyteczności, ma subiektywne regularności, gdy jest rozpatrywane jako problem poznawczy, oraz zewnętrzne prawidłowości, gdy jest rozpatrywane jako struktura zdarzeń. Te prawidłowości nazywane są podcelami. Podcele definiują obszary, w których problem można skutecznie rozpatrywać z lokalnej perspektywy. Zamiast konieczności przemyślenia przez umysł całego łańcucha rozumowania „Działanie A prowadzi do B, które prowadzi do C, które prowadzi do D, ..., które prowadzi do rzeczywistej wewnętrznej użyteczności Z", istnieje użyteczna prawidłowość, że działania prowadzące do B są w większości przewidywane, aby prowadzić przez łańcuch do Z. Podobnie umysł może rozważyć, które z podcelów B1, B2, B3 najprawdopodobniej doprowadzą do C lub rozważyć, które podcele C1, C2, C3 są łącznie wystarczające dla D, bez ponownego przemyślenia reszty logiki do Z. Ta struktura zdarzeń sieciowych (nie hierarchicznych) jest niedoskonałą prawidłowością; pożądanie jest dziedziczne tylko w takim zakresie i dokładnie w takim zakresie, w jakim przewidywane prowadzenie do Z jest dziedziczne. Nasz wszechświat o niskiej entropii ma strukturę kategorii, ale nie doskonałą strukturę kategorii. Używanie obrazowania do opisu zdarzenia E, które ma prowadzić do zdarzenia F, nigdy nie jest idealne; być może większość stanów świata rzeczywistego, które pasują do opisu E, prowadzi do zdarzeń, które pasują do opisu F, ale poza czystą matematyką bardzo rzadko można znaleźć przypadek, w którym przewidywanie jest idealne. Zawsze będą pewne stany w objętości wyodrębnionej przez opis E, które prowadzą do stanów poza objętością wyodrębnioną przez opis F. Jeśli przewiduje się, że C prowadzi do D, a B przewiduje się, że prowadzi do C, to zazwyczaj B odziedziczy przewidywaną do-prowadzenia-do-D-owość C. Może się jednak zdarzyć, że B prowadzi do szczególnego przypadku C, który nie prowadzi do D; w takim przypadku B nie odziedziczy przewidywanej do-prowadzenia-do-D-owości C. Dlatego też, gdyby C odziedziczył pożądanie od D, B również nie odziedziczyłby pożądania C. Aby poradzić sobie ze światem niedoskonałych regularności, systemy celów modelują regularności w nieregularnościach, używając ograniczeń opisowych, odległych splątań i globalnych heurystyk. Jeśli zdarzenia pasujące do opisu E zwykle, ale nie zawsze, prowadzą do zdarzeń pasujących do opisu F, wówczas wyobrażenia mentalne opisujące E, a nawet koncepcje tworzące opis E, mogą zostać udoskonalone w celu zawężenia klasy ekstensjonalnej w celu wyeliminowania zdarzeń, które wydają się pasować do E, ale nie okazują się prowadzić do F. Te „ograniczenia opisowe" sprawiają, że sztuczna inteligencja skupia się na koncepcjach i kategoriach, które ujawniają predykcyjne, przyczynowe i manipulowalne regularności w rzeczywistości, a nie tylko regularności powierzchniowe. Kolejnym udoskonaleniem są „odległe splątania"; na przykład działanie A, które prowadzi do B, które prowadzi do C, ale które jednocześnie ma skutki uboczne, które blokują D, co jest źródłem pożądania C. Innym rodzajem splątania jest sytuacja, gdy działanie A prowadzi do niezwiązanego efektu ubocznego S, który ma negatywną użyteczność przeważającą nad pożądanością odziedziczoną po B. „Globalne heurystyki" opisują regularności celów, które są ogólne w wielu kontekstach problemu i które mogą być zatem używane do szybkiego rozpoznawania pozytywnych i negatywnych cech; koncepcja „margines błędu" jest kategorią opisującą ważną cechę wielu planów, a przekonanie, że „margines błędu wspiera lokalny cel" jest globalną heurystyką, która pozytywnie łączy członków kategorii percepcyjnej margines błędu z lokalnym kontekstem celu, bez konieczności oddzielnego podsumowania indukcyjnego i dedukcyjnego wsparcia dla ogólnej heurystyki. Podobnie w samomodyfikujących się lub przynajmniej samoregulujących SI „minimalizacja wykorzystania pamięci" jest podcelem, na który może mieć wpływ wiele innych podcelów i działań, więc percepcyjne rozpoznanie zdarzeń w kategorii „użycie pamięci" lub „prowadzi do wykorzystania pamięci" oznacza splątanie z określonym odległym celem. Ograniczenia opisowe, odległe splątania i globalne heurystyki nie naruszają modelu pożądalności jako przewidywania; ograniczenia opisowe, odległe splątania i globalne heurystyki są również przydatne do modelowania złożonych przewidywań, w ten sam sposób i z tych samych powodów, z których są przydatne do modelowania celów. Istnieją jednak co najmniej trzy powody, dla których aktywność planowania różni się od aktywności przewidywania. Po pierwsze, przewidywanie zazwyczaj postępuje naprzód od określonego stanu wszechświata, aby określić, co

nastąpi później, podczas gdy planowanie często (choć nie zawsze) rozumuje wstecz od wyobrażeń celu, aby wybrać jeden punkt w przestrzeni możliwych wszechświatów, przy czym wymiary przestrzeni są określone przez stopnie swobody w dostępnych działaniach. Po drugie, pożądalności są różniczkowe, w przeciwieństwie do przewidywań; jeśli A i ∼ A prowadzą do tego samego punktu końcowego E, to z punktu widzenia predykcyjnego może to zwiększyć zaufanie do E, ale z punktu widzenia planowania oznacza to, że ani A ani ∼ A nie odziedziczą czystej pożądaności od E. Końcowym efektem pożądalności jest to, że AI wybiera najbardziej pożądane działanie, operację, która jest porównawcza, a nie absolutna; jeśli zarówno A, jak i ∼ A prowadzą do E, ani A, ani ∼ A nie przekazują różnicowej pożądalności działaniom. Po trzecie, podczas gdy zarówno implikacja, jak i związek przyczynowy są przydatne do rozumowania o przewidywaniach, tylko związki przyczynowe są przydatne w rozumowaniu o celach. Jeśli po obserwacji A zwykle następuje obserwacja B, to czyni to A dobrym predyktorem B – niezależnie od tego, czy A jest bezpośrednią przyczyną B, czy też istnieje ukryta trzecia przyczyna C, która jest bezpośrednią przyczyną zarówno A, jak i B. Uważałbym implikację za wyłaniającą się właściwość ukierunkowanej sieci zdarzeń, której podstawowe zachowanie jest zachowaniem związku przyczynowego; jeśli C powoduje A, a następnie powoduje B, to A będzie implikować B. Zarówno „A powoduje B" (bezpośredni związek przyczynowy), jak i „A implikuje B" (wzajemny związek przyczynowy z C) są przydatne w przewidywaniu. Jednak w planowaniu rozróżnienie między „A bezpośrednio powoduje B" a „A i B są efektami C" prowadzi do rozróżnienia między „Działania, które prowadzą do A, jako takie, prawdopodobnie doprowadzą do B" a „Działania, które prowadzą bezpośrednio do A, bez wcześniejszego prowadzenia przez C, prawdopodobnie nie będą miały żadnego wpływu na B". To rozróżnienie oznacza również, że eksperymenty w manipulacji mają tendencję do wyodrębniania rzeczywistych powiązań przyczynowych w sposób, w jaki nie robią tego testy predykcyjne. Jeśli A implikuje B, to często zdarza się, że C powoduje zarówno A, jak i B, ale w większości problemów ze świata rzeczywistego rzadziej zdarza się, aby działanie mające na celu wpłynięcie na A oddzielnie i niewidocznie wpłynęło na ukrytą trzecią przyczynę C, co prowadzi do fałszywego potwierdzenia bezpośredniej przyczynowości. (Chociaż zdarza się to, szczególnie w eksperymentach ekonomicznych i psychologicznych.)

Działania inteligencji: Wyjaśnienie, przewidywanie, odkrywanie, planowanie, projektowanie

Do tej pory w tej sekcji wprowadzono rozróżnienie między modelami sensorycznymi, predykcyjnymi, decyzyjnymi i manipulacyjnymi; zmiennymi dyskretnymi, ilościowymi i wzorcowanymi; holonicznym modelem wzorców wysokiego i niskiego poziomu; oraz odniesieniami supercelów, obrazowaniem celów i działaniami. Te idee zapewniają ramy do zrozumienia bezpośrednich podzadań inteligencji – chwilowych działań deliberacyjnych. Podczas wykonywania zadania poznawczego wysokiego poziomu, takiego jak zaprojektowanie roweru, podzadania składają się z przekraczania luk od bardzo wysokiego poziomu holonów, takich jak dobry transport, do holonu szybkiego napędu, do holonu pchającego na ziemi, do holonu koła, do holonów szprych i opon, aż w końcu holony staną się bezpośrednio możliwe do określenia pod względem komponentów projektowych i materiałów projektowych bezpośrednio dostępnych dla SI. Działania inteligencji można opisać jako uzupełnianie wiedzy w celu realizacji celu. Aby ukończyć rower, najpierw należy ukończyć projekt roweru. Aby zrealizować plan, należy ukończyć mentalny obraz planu. Ponieważ zarówno planowanie, jak i projektowanie intensywnie wykorzystują wiedzę, często rodzą one działania czysto zorientowane na wiedzę, takie jak wyjaśnianie, przewidywanie i odkrywanie. Działania te są chaotycznymi, nieinkluzywnymi kategoriami, ale ilustrują ogólne rodzaje rzeczy, które robią ogólne umysły. Działania związane z wiedzą są wykonywane zarówno na dużą skalę, jako główne cele strategiczne, jak i na małą skalę, w rutynowych podzadaniach. Na przykład „wyjaśnienie" dąży do rozszerzenia bieżącej wiedzy, poprzez dedukcję, indukcję lub eksperyment, aby wypełnić lukę pozostawioną przez nieznaną przyczynę znanego skutku. Nieznana przyczyna będzie przynajmniej punktem odniesienia dla obrazów pytań, które wprowadzą do gry

sekwencje i weryfikatory, które reagują na pytania otwarte. Jeśli problem stanie się wystarczająco wyrazisty i trudny, znalezienie nieznanej przyczyny może zostać przeniesione z obrazów pytań do wewnętrznego celu, umożliwiając SI rozważne rozumowanie na temat tego, jakie strategie rozwiązywania problemów wdrożyć. Cel wiedzy dotyczący „budowania planu" dziedziczy pożądanie z celu planu, ponieważ stworzenie planu jest wymagane do (jest podcelem) osiągnięcia celu planu. Cel wiedzy dotyczący wyjaśnienia zaobserwowanej awarii może dziedziczyć pożądanie z celu możliwego do osiągnięcia po naprawieniu awarii. Ponieważ cele wiedzy mogą rządzić rzeczywistymi działaniami, a nie tylko przepływem następstw, należy je odróżnić od obrazowania pytań. Cele wiedzy umożliwiają również refleksyjne rozumowanie na temat tego, jakie działania wewnętrzne prawdopodobnie doprowadzą do rozwiązania problemu; cele wiedzy mogą wywoływać następstwa, które poszukują przekonań na temat rozwiązywania problemów wiedzy, a nie tylko przekonań na temat konkretnego problemu. Wyjaśnienie wypełnia luki w wiedzy o przeszłości. Przewidywanie wypełnia luki w wiedzy o przyszłości. Odkrycie wypełnia luki w wiedzy o teraźniejszości. Projekt wypełnia luki w modelu mentalnym narzędzia. Planowanie wypełnia luki w modelu przyszłych strategii i działań. Wyjaśnienie, przewidywanie, odkrywanie i projektowanie mogą być wykorzystywane w dążeniu do określonego celu w świecie rzeczywistym lub jako niezależne dążenie w oczekiwaniu na przydatność uzyskanej wiedzy w przyszłych celach – „ciekawość". Ciekawość całkowicie wypełnia ogólne luki (zamiast być ukierunkowaną na konkretne, już znane luki) i obejmuje wykorzystanie rozumowania i eksperymentowania zorientowanego na przyszłość, zamiast wstecznego łączenia się ze specyficznymi pożądanymi celami wiedzy; ciekawość można postrzegać jako wypełnianie bardzo abstrakcyjnego celu „odkrycia X, gdzie X odnosi się do czegokolwiek, co okaże się dobrą rzeczą do poznania później, nawet jeśli nie wiem dokładnie, czym jest X". (Ciekawość obejmuje bardzo abstrakcyjne powiązanie z wewnętrzną użytecznością, ale takie, które jest mimo wszystko całkowicie prawdziwe – ciekawość jest przydatna). Wszystkie te działania mają to do siebie, że obejmują rozumowanie na temat złożonego, holonicznego modelu przyczyn i skutków. „Wyjaśnienie" wypełnia luki dotyczące przeszłości, która jest złożonym systemem przyczyn i skutków. „Przewidywanie" wypełnia luki w przyszłości, która jest złożonym systemem przyczyn i skutków. „Projektowanie" rozumuje o narzędziach, które są złożonymi holonicznymi systemami przyczyn i skutków. „Planowanie" rozumuje o strategiach, które są złożonymi holonicznymi systemami przyczyn i skutków. Inteligentne rozumowanie uzupełnia cele wiedzy i odpowiada na pytania w złożonym holonicznym modelu przyczynowym, aby osiągnąć odniesienia celów w złożonym holonicznym systemie przyczynowym. Daje nam to trzy elementy DGI:

• Czym jest inteligencja: Inteligencja składa się u ludzi z wysoce modułowego mózgu z dziesiątkami obszarów, który wdraża proces deliberatywny (zbudowany na myślach zbudowanych z koncepcji zbudowanych na modalnościach sensorycznych zbudowanych na neuronach); plus podsystemy składowe (np. pamięć); plus otaczające podsystemy (np. regulacja autonomiczna); plus pozostałe podsystemy wdrażające przeddeliberatywne przybliżenia procesów deliberatywnych; plus emocje, instynkty, intuicje i inne systemy, które wpływają na proces deliberatywny w sposób, który był adaptacyjny w środowisku przodków; plus wszystko inne. Podobny system jest rozważany dla SI, mniej więcej tego samego rzędu złożoności, ale nieuchronnie mniej chaotyczny. Oba supersystemy charakteryzują się poziomami organizacji: kodem / neuronami, modalnościami, koncepcjami, myślami i deliberacją.

• Dlaczego jest inteligencja: Przyczyną ludzkiej inteligencji jest ewolucja. Inteligencja jest ewolucyjną zaletą, ponieważ umożliwia nam modelowanie rzeczywistości, w tym rzeczywistości zewnętrznej, rzeczywistości społecznej i rzeczywistości wewnętrznej, co z kolei umożliwia nam przewidywanie, decydowanie i manipulowanie rzeczywistością. Sztuczna inteligencja będzie inteligentna, ponieważ my, ludzcy programiści, chcemy osiągnąć cel, który najlepiej można osiągnąć za pomocą inteligentnej sztucznej inteligencji, lub ponieważ uważamy, że akt tworzenia sztucznej inteligencji ma wewnętrzną

użyteczność; w obu przypadkach budowanie sztucznej inteligencji wymaga zbudowania przemyślanego supersystemu, który manipuluje rzeczywistością.

• Jak inteligencja: Inteligencja (świadome rozumowanie) realizuje cele wiedzy i odpowiada na pytania w złożonym holonicznym modelu przyczynowym, aby osiągnąć odniesienia celów w złożonym holonicznym systemie przyczynowym.

Ogólna inteligencja

Kontekst ewolucyjny inteligencji historycznie obejmował środowiskowe konteksty adaptacyjne, społeczne konteksty adaptacyjne (modelowanie innych umysłów) i refleksyjne konteksty adaptacyjne (modelowanie wewnętrznej rzeczywistości). Ewoluując, aby dopasować się do szerokiej gamy kontekstów adaptacyjnych, nabyliśmy wiele funkcji poznawczych, które są widocznie wyspecjalizowane w określonych problemach adaptacyjnych, ale nabyliśmy również funkcje poznawcze, które są adaptacyjne w wielu kontekstach, oraz funkcje adaptacyjne, które kooptują wcześniej wyspecjalizowane funkcje do szerszego zastosowania. Ludzie mogą nabyć znaczną kompetencję w modelowaniu, przewidywaniu i manipulowaniu w pełni ogólnymi prawidłowościami naszego wszechświata o niskiej entropii. Nazywamy tę zdolność „ogólną inteligencją". W pewnym sensie nasza zdolność jest bardzo słaba; często rozwiązujemy ogólne problemy abstrakcyjnie, a nie percepcyjnie, więc nie możemy rozwiązywać problemów celowo w kolejności wizualnej interpretacji sceny 3D w czasie rzeczywistym. Ale często możemy powiedzieć coś, co jest wystarczająco prawdziwe, aby było przydatne i wystarczająco proste, aby było wykonalne. Możemy rozważać, jak działa widzenie, nawet jeśli nie potrafimy rozważać wystarczająco szybko, aby wykonywać przetwarzanie wizualne w czasie rzeczywistym. Obecnie istnieje szeroki trend w kierunku mapowania jeden do jednego podsystemów poznawczych na kompetencje domenowe. Przyznaję, że osobiście denerwują mnie przejawy tej idei w popularnej psychologii, ale oczywiście nowe frenologie są nieistotne dla prawdziwych hipotez dotyczących mapowania między wyspecjalizowanymi kompetencjami domenowymi a wyspecjalizowanymi podsystemami obliczeniowymi lub decyzji o dążeniu do wyspecjalizowanej sztucznej inteligencji. Nie jest niczym niezwykłym, że trendy akademickie odzwierciedlają popularną psychologię, ale ogólnie rzecz biorąc, dobrą formą jest pozbycie się tezy przed analizą wad moralnych jej zwolenników. W DGI uważa się, że ludzka inteligencja składa się z supersystemu ze złożonymi, współzależnymi podsystemami, które wykazują wewnętrzną specjalizację funkcjonalną, ale nie wyklucza to istnienia innych podsystemów, które przyczyniają się wyłącznie lub głównie do określonych talentów poznawczych i kompetencji domenowych, lub podsystemów, które przyczyniają się w większym stopniu do niektórych talentów poznawczych niż inne. Mapowanie z podsystemów obliczeniowych na talenty poznawcze jest wielorakie, a mapowanie z talentów poznawczych plus nabytej wiedzy eksperckiej na kompetencje domenowe jest również wielorakie, ale nie wyklucza to konkretnych odpowiedników między ludzkimi wariancjami w „mocy obliczeniowej" (uogólnione zasoby poznawcze) przydzielonej do podsystemów obliczeniowych a obserwowanymi wariancjami w talentach poznawczych lub kompetencjach domenowych. Należy jednak zauważyć, że przedmiotem AI nie jest wariancja między ludźmi, ale podstawa adaptacyjnej złożoności wspólna dla wszystkich ludzi. Jeśli zwiększenie zasobów przydzielonych do podsystemu poznawczego powoduje wzrost talentu poznawczego lub kompetencji domenowej, nie wynika z tego, że talent lub kompetencja mogą być implementowane przez sam ten podsystem. Należy również zauważyć, że zgodnie z tradycyjnym paradygmatem programowania, myśli programistów na temat rozwiązywania konkretnych problemów są tłumaczone na kod, a jest to idiom leżący u podstaw większości gałęzi klasycznej AI; na przykład inżynierowie systemów eksperckich rzekomo tłumaczą przekonania w określonych domenach bezpośrednio na treść poznawczą SI. To naturalnie prowadziłoby do wizji inteligencji, w której istnieje bezpośrednie odwzorowanie między podsystemami i kompetencjami.

Wierzę, że jest to podstawowa przyczyna atmosfery, w której poszukiwanie inteligentnej SI jest witane odpowiedzią: „SI, która jest inteligentna w jakiej domenie?". Nie oznacza to, że eksploracja wyspecjalizowanej SI jest całkowicie bezwartościowa; w rzeczywistości poziomy organizacji DGI sugerują określoną klasę przypadków, w których wyspecjalizowana SI może okazać się owocna. Modalności sensoryczne znajdują się bezpośrednio nad poziomem kodu; modalności sensoryczne były jednymi z pierwszych wyspecjalizowanych podsystemów poznawczych, które ewoluowały, a zatem nie są tak zależne od wspierających ram supersystemu, chociaż inne części supersystemu w dużym stopniu zależą od modalności. Sugeruje to, że wyspecjalizowane podejście, w którym programiści bezpośrednio piszą kod, może okazać się owocne, jeśli projekt konstruuje modalność sensoryczną. I rzeczywiście, badania nad SI, które koncentrują się na tworzeniu systemów sensorycznych i systemów sensomotorycznych, nadal przynoszą rzeczywisty postęp. Tacy badacze podążają ścieżką przyrostową ewolucji, często świadomie, unikając w ten sposób pułapek wynikających z naruszania poziomów organizacji. Nadal jednak nie wierzę, że możliwe jest dopasowanie szerokiej stosowalności deliberatywnego supersystemu poprzez wdrożenie oddzielnego podsystemu obliczeniowego dla każdego kontekstu problemu. Nie tylko niemożliwe jest zduplikowanie ogólnej inteligencji poprzez sumę takich podsystemów, ale podejrzewam, że niemożliwe jest osiągnięcie wydajności na poziomie ludzkim w większości pojedynczych kontekstów przy użyciu wyspecjalizowanej sztucznej inteligencji. Czasami używamy abstrakcyjnej deliberacji, aby rozwiązać problemy na poziomie modalności, dla których brakuje nam modalności sensorycznych, i w tym przypadku projekty AI mogą rozwiązać problem na poziomie modalności, ale wynikająca z tego metoda rozwiązywania problemów będzie bardzo różna od ludzkiej i nie będzie uogólniać się poza konkretną domenę. Stąd Deep Blue. Nawet na poziomie indywidualnych kompetencji domenowych nie wszystkie kompetencje są ze sobą niepowiązane. Różne umysły mogą mieć różne zdolności w różnych domenach; umysł może mieć „powierzchnię zdolności" z pagórkami i szczytami w obszarach wysokiej zdolności; ale szczyt w obszarze takim jak uczenie się lub samodoskonalenie ma tendencję do podnoszenia pozostałej powierzchni zdolności [103]. Talenty i podsystemy, które są ogólne w sensie przyczyniania się do wielu kompetencji domenowych – i kompetencji domenowych samodoskonalenia; patrz sekcja 3 – zajmują strategiczną pozycję w AI analogicznie do centralnych pól w szachach.

Ja

Kiedy AI może legalnie używać słowa „ja"? (Na potrzeby tej dyskusji muszę nadać AI tymczasową nazwę własną; podczas tej dyskusji będę używać „Aisa"). Klasyczna AI, która zawiera token LISP dla „hamburgera", nie wie nic o hamburgerach; co najwyżej AI może rozpoznawać powtarzające się wystąpienia ciągu liter wpisywanych przez programistów. Nadanie AI sugestywnie nazwanej struktury danych lub funkcji nie czyni tego komponentu funkcjonalnym odpowiednikiem podobnie nazwanej cechy ludzkiej [66]. W którym momencie Aisa może mówić o czymś zwanym „Aisa", aby Drew McDermott nie wyskoczył i nie oskarżył nas o używanie terminu, który równie dobrze można by przetłumaczyć jako „G0025"? Załóżmy, że Aisa, oprócz modelowania środowisk wirtualnych i/lub świata zewnętrznego, modeluje również pewne aspekty rzeczywistości wewnętrznej, takie jak skuteczność przekonań heurystycznych używanych w różnych sytuacjach. Stopnie powiązania między modelem a rzeczywistością są sensoryczne, predykcyjne, decydujące i manipulacyjne. Załóżmy, że Aisa potrafi wyczuć, kiedy heurystyka jest stosowana, zauważyć, że heurystyki mają tendencję do stosowania w pewnych kontekstach i że mają tendencję do dawania pewnych rezultatów, i wykorzystać ten indukcyjny dowód do sformułowania oczekiwań co do tego, kiedy heurystyka zostanie zastosowana i przewidzieć rezultaty jej zastosowania. Aisa teraz predykcyjnie modeluje Aisę; tworzy przekonania na temat jej działania, obserwując introspektywnie widoczne efekty jej podstawowych mechanizmów. Zacieśnienie powiązania z predyktywnego na manipulacyjne wymaga, aby Aisa powiązała introspekcyjne obserwacje z wewnętrznymi działaniami; na przykład Aisa może zauważyć,

że poświęcenie dyskrecjonalnej mocy obliczeniowej pewnemu podprocesowi daje myśli pewnego rodzaju i że myśli tego rodzaju są przydatne w pewnych kontekstach, a następnie poświęcić dyskrecjonalną moc temu podprocesowi w tych kontekstach. Manipulacyjne powiązanie między Aisą a modelem Aisy Aisy wystarczy, aby Aisa mogła zasadnie powiedzieć „Aisa używa heurystyki X", tak że użycie terminu „Aisa" jest istotnie różne od użycia „hamburgera" lub „G0025". Ale czy Aisa może zasadnie powiedzieć „Używam heurystyki X"? Mój ulubiony cytat na ten temat pochodzi od Douglasa Lenata, chociaż nie mogę znaleźć odniesienia i cytuję z pamięci: „Podczas gdy Cyc wie, że istnieje coś zwanego Cyc i że Cyc jest komputerem, nie wie, że jest Cyc."33 Osobiście kwestionowałbym, czy Cyc wie, że Cyc jest komputerem – ale niezależnie od tego Lenat dokonał uzasadnionego i fundamentalnego rozróżnienia. Modelowanie przez Aisę czegoś zwanego Aisą nie jest tym samym, co modelowanie przez Aisę samej siebie. W dziwnym sensie założenie, że problem istnieje, wystarczy, aby rozwiązać problem. Jeśli wymagany jest kolejny krok, zanim Aisa będzie mogła powiedzieć „Używam heurystyki X", to musi istnieć istotna różnica między stwierdzeniem „Aisa używa heurystyki X" a „Używam heurystyki X". I to jest jedna z możliwych odpowiedzi: Aisa może powiedzieć „ja", gdy zachowanie samego modelowania jest istotnie inne, ze względu na samoodniesienie, od zachowania modelowania innej sztucznej inteligencji, która wygląda jak Aisa. Jednym konkretnym przypadkiem, w którym samomodelowanie jest istotnie inne niż inne modelowanie, jest planowanie. Zastosowanie złożonego planu, w którym liniowa sekwencja działań A, B, C jest indywidualnie konieczna i łącznie wystarczająca do osiągnięcia celu G, wymaga domniemanego założenia, że sztuczna inteligencja zrealizuje własne plany; działanie A jest bezużyteczne, chyba że nastąpią po nim działania B i C, a działanie A nie jest zatem pożądane, chyba że przewiduje się, że nastąpią po nim działania B i C. Tworzenie złożonych planów nie wymaga w rzeczywistości samomodelowania, ponieważ wiele klasycznych SI angażuje się w zachowania podobne do planowania, używając programowych założeń zamiast refleksyjnego rozumowania, a u ludzi założenie to jest zwykle automatyczne, a nie przedmiotem rozważań. Jednak celowe refleksyjne rozumowanie dotyczące złożonych planów wymaga zrozumienia, że przyszłe działania SI są determinowane przez decyzje przyszłego ja SI, że istnieje pewien stopień ciągłości (choć nie jest to ciągłość doskonała) między obecnym a przyszłym ja, a zatem istnieje pewien stopień ciągłości między obecnymi decyzjami a przyszłymi działaniami. Inteligentny umysł porusza się po wszechświecie z czterema głównymi klasami zmiennych: czynniki losowe, zmienne z ukrytymi wartościami, działania innych agentów i działania samego siebie. Przestrzeń możliwych działań różni się od przestrzeni wydzielonych przez inne zmienne, ponieważ przestrzeń możliwych działań jest pod kontrolą SI. Jedną różnicą między „Aisa użyje heurystyki X" a „Użyję heurystyki X" jest stopień, w jakim użycie heurystyki jest pod świadomą kontrolą Aisy – stopień, w jakim Aisa ma cele związane z użyciem heurystyki, a zatem stopień, w jakim obserwacja „Przewiduję, że użyję heurystyki X" wpływa na późniejsze działania Aisy. Aisa, jeśli jest wystarczająco kompetentna w modelowaniu innych umysłów, mogłaby przewidzieć, że podobna SI o imieniu Aileen również użyje heurystyki X, ale przekonania na temat zachowań Aileen wynikałyby z predykcyjnego modelowania Aileen, a nie z decydującego planowania wewnętrznych działań w oparciu o celowy wybór z przestrzeni możliwości. Istnieje różnica poznawcza między stwierdzeniem Aisy „Przewiduję, że Aileen użyje heurystyki X" a „Planuję użyć heurystyki X". Na poziomie systemowym globalna wyjątkowość „ja" byłaby przybita tymi heurystykami, przekonaniami i oczekiwaniami, które indywidualnie odnoszą się specjalnie do „ja" z powodu introspektywnej refleksyjności lub przestrzeni niezdecydowanych, ale decyzyjnych działań. Moim zdaniem taka sztuczna inteligencja byłaby w stanie legalnie używać słowa „ja", chociaż u ludzi wyjątkowość „ja" może być również przybita przez dodatkowe siły poznawcze. (Uzasadnione użycie „ja" nie jest wyraźnie oferowane jako konieczny i wystarczający warunek dla „trudnego problemu świadomego doświadczenia" lub społecznej, prawnej i moralnej osobowości).

Seed AI

W przestrzeni pomiędzy teorią inteligencji ludzkiej a teorią ogólnej SI znajduje się widmowy zarys teorii umysłów w ogóle, wyspecjalizowanej dla ludzi i SI. Nie próbowałem przedstawić takiej teorii wprost, ograniczając się do omówienia tych konkretnych podobieństw i różnic między ludźmi a SI, które moim zdaniem warto zgadnąć z góry. Rewolucja kopernikańska w naukach kognitywnych – ludzie jako niecentralny przypadek szczególny – nie jest jeszcze gotowa; potrzeba trzech punktów, aby narysować krzywą, a obecnie mamy tylko jeden. Niemniej jednak ludzie są w rzeczywistości niecentralnym przypadkiem szczególnym, a ten abstrakcyjny fakt jest poznawalny, nawet jeśli nasze obecne teorie są antropocentryczne. Istnieje fundamentalna przepaść między ewolucyjnym projektem a projektem deliberatywnym. Z perspektywy inteligencji deliberatywnej – na przykład człowieka – ewolucja jest zdegenerowanym przypadkiem projektowania i testowania, w którym inteligencja równa się zeru. Mutacje są atomowe; rekombinacje są losowe; zmiany są dokonywane na najniższym poziomie organizacji genotypu (odwracanie bitów genetycznych); wielkość ziarna testowanego składnika to cały organizm; a metryka dobroci działa wyłącznie poprzez indukcję w historycznie napotkanych przypadkach, bez dedukcyjnego rozumowania o tym, które czynniki kontekstowe mogą się później zmienić. Ewolucja ewolucyjności nieco poprawia ten obraz. Istnieje tendencja, aby niskopoziomowe bity genetyczne sprawowały kontrolę nad złożonością wysokiego poziomu, tak że zmiany w tych genach mogą tworzyć zmiany wysokiego poziomu. Ślepe naciski selekcyjne mogą tworzyć samonaprawiające się i samoprzewodzące systemy, które okazują się wysoce ewolucyjne ze względu na ich zdolność do fenotypowego dostosowywania się do zmian genotypowych. Niemniej jednak ewolucja ewolucyjności nie jest substytutem inteligentnego projektu. Ewolucja działa, pomimo lokalnych nieefektywności, ponieważ ewolucja wywiera ogromną kumulatywną presję projektową w czasie. Jednak całkowita ilość nacisku projektowego wywieranego w danym czasie jest ograniczona; istnieje tylko ograniczona ilość nacisku selekcyjnego, który należy podzielić między wszystkie wariancje genetyczne wybrane w danym pokoleniu. Jedną oczywistą konsekwencją jest to, że ewolucyjnie niedawne adaptacje będą prawdopodobnie mniej zoptymalizowane niż te, które są ewolucyjnie starożytne. W DGI ewolucyjna filogeneza inteligencji mniej więcej podsumowuje jej funkcjonalną ontogenezę; wynika z tego, że wyższe poziomy organizacji mogą zawierać mniej całkowitej złożoności niż niższe poziomy, chociaż czasami wyższe poziomy organizacji są również bardziej ewolucyjne. Dlatego subtelniejszą konsekwencją jest to, że niższe poziomy organizacji prawdopodobnie będą mniej dobrze dostosowane do ewolucyjnie niedawnych innowacji (takich jak rozważanie) niż te wyższe poziomy do niższych poziomów – efekt wzmocniony przez właściwości ewolucji zachowujące strukturę, w tym zachowanie struktury, która ewoluowała w nieobecności rozważania. Wszelkie możliwości projektowe, które po raz pierwszy otworzyły się wraz z pojawieniem się Homo sapiens sapiens, pozostają niewykorzystane, ponieważ Homo sapiens sapiens istnieje zaledwie od 50 000–100 000 lat; jest to wystarczająco dużo czasu, aby dokonać wyboru spośród odchyleń w tendencjach ilościowych, ale nie jest to wystarczająco dużo czasu, aby skonstruować złożoną adaptację funkcjonalną. Ponieważ wiadomo, że tylko Homo sapiens sapiens w swojej najnowocześniejszej formie zajmuje się programowaniem komputerowym, może to wyjaśniać, dlaczego nie mamy jeszcze zdolności do przeprogramowania własnych neuronów (mówię to z przymrużeniem oka, ale jest w tym ziarno prawdy). A ewolucja jest niezwykle konserwatywna, jeśli chodzi o całkowitą rewizję architektur; geny homeotyczne kontrolujące różnicowanie embrionalne przedniego, śródmózgowia i tylnego mózgu mają identyfikowalne homologi w rozwijającej się głowie muchy Drosophila (!). Ewolucja nigdy nie refaktoryzuje swojego kodu. Znacznie łatwiej jest ewolucji potknąć się o tysiąc indywidualnych optymalizacji niż o dwie równoczesne zmiany, które są razem korzystne i osobno szkodliwe. Kod genetyczny, który określa mapowanie między kodonami (kodon to trzy zasady DNA) i 20 aminokwasami, jest nieefektywny; mapuje 64 możliwe kodony na 20 aminokwasów plus kod stop. Dlaczego ewolucja nie przesunęła jednego z obecnie zbędnych kodonów do nowego aminokwasu, rozszerzając tym samym zakres możliwych białek? Ponieważ dla każdego złożonego organizmu

najmniejsza zmiana w zachowaniu DNA – najniższy poziom organizacji genetycznej – zniszczyłaby praktycznie wszystkie wyższe poziomy adaptacyjnej złożoności, chyba że zmianie towarzyszyłyby miliony innych jednoczesnych zmian w całym genomie, aby przesunąć każdy nagle niestandardowy kodon do jednego z jego poprzednich odpowiedników. Ewolucja po prostu nie jest w stanie poradzić sobie z jednoczesnymi zależnościami, chyba że poszczególne zmiany można wdrażać stopniowo lub że w wyniku pojedynczej zmiany genetycznej wystąpi wiele efektów fenotypowych. Dla ludzi planowanie skoordynowanych zmian jest rutyną; dla ewolucji jest niemożliwe. Ewolucja jest uderzana ogromną stopą dyskontową przy wymianie papierowej waluty przyrostowej optymalizacji na twardą monetę złożonego projektu. Powinniśmy oczekiwać, że ludzki projekt będzie zawierał przerażająco dużą liczbę prostych optymalizacji funkcjonalnych. Ale jest również zrozumiałe, jeśli istnieją deficyty w wyższym projekcie. Podczas gdy wyższe poziomy organizacji (w tym rozważania) wyłoniły się z niższych poziomów i stąd są dość dobrze do nich dostosowane, niższe poziomy organizacji nie są tak dostosowane do istnienia świadomej inteligencji. Ludzie zostali skonstruowani przez akrecyjne procesy ewolucyjne, przechodząc od bardzo złożonej inteligencji nieogólnej do bardzo złożonej inteligencji ogólnej, przy czym rozważanie jest ostatnią warstwą wisienki na torcie. Czy możemy wymienić twardą monetę złożonego projektu na papierową walutę optymalizacji niskiego poziomu? „Optymalizacja kompilatorów" to oczywisty krok, ale niewielki; optymalizacja programu przyspiesza programy, ale nie wywiera presji projektowej na lepszą organizację funkcjonalną, nawet w przypadku prostych funkcji tego rodzaju, które łatwo optymalizuje ewolucja. Kierowana ewolucja, stosowana w przypadku modułowych podzadań z jasno zdefiniowanymi metrykami wydajności, byłaby nieco większym krokiem. Ale nawet ukierunkowana ewolucja jest nadal zdegenerowanym przypadkiem projektowania i testowania, w którym poszczególne kroki są nieinteligentne. Z założenia budujemy sztuczną inteligencję. Po co stosować nieinteligentne projektowanie i testowanie? Niewątpliwie istnieje granica typu „co było pierwsze — jajko czy kura" w poleganiu na inteligencji sztucznej inteligencji w budowaniu sztucznej inteligencji. Dopóki nie zostanie osiągnięty stabilnie funkcjonujący supersystem poznawczy, dostępna będzie tylko nierozważna inteligencja wykazywana przez części systemu. Nawet po osiągnięciu funkcjonującego supersystemu — co samo w sobie jest bohaterskim wyczynem — inteligencja wykazywana przez ten supersystem będzie początkowo bardzo słaba. Im słabsza inteligencja sztucznej inteligencji, tym mniejsze zdolności wykaże w rozumieniu złożonych systemów holonicznych. Im słabsze zdolności sztucznej inteligencji w zakresie projektowania holonicznego, tym mniejsze części samej siebie będzie w stanie zrozumieć. Kiedykolwiek sztuczna inteligencja w końcu stanie się wystarczająco inteligentna, aby uczestniczyć we własnym tworzeniu, początkowo będzie musiała skoncentrować się na ulepszaniu małych części samej siebie za pomocą prostych i jasnych metryk wydajności dostarczanych przez programistów. To nie jest szczególny przypadek głupiej SI próbującej zrozumieć samą siebie, ale szczególny przypadek głupiej SI próbującej zrozumieć dowolny złożony system holoniczny; gdy SI jest „młoda", prawdopodobnie będzie ograniczona do rozumienia prostych elementów systemu lub małych organizacji elementów i tylko tam, gdzie istnieją jasne konteksty celów (prawdopodobnie wyjaśnione przez programistę). Ale nawet prymitywna zdolność projektowania holonicznego mogłaby wypełnić lukę ludzką; nie lubimy bawić się drobiazgami, ponieważ nudzimy się i nie mamy możliwości zamiany naszej ogromnej mocy równoległej w przypadku złożonych problemów na większą prędkość szeregową w przypadku prostych problemów. Podobnie byłoby niezdrowe (skutkowałoby patologiami SI), gdyby ludzkie zdolności programistyczne odgrywały stałą rolę w uczeniu się lub optymalizacji jąder koncepcji – ale w punktach, w których ingerencja wydaje się kusząca, jest całkowicie dopuszczalne, aby procesy deliberatywne SI odgrywały rolę, jeśli SI posunęła się tak daleko. Ludzka inteligencja, stworzona przez ewolucję, charakteryzuje się sygnaturą projektu ewolucji. Zdecydowana większość naszej historii genetycznej miała miejsce przy braku inteligencji deliberatywnej; nasze starsze systemy poznawcze są słabo przystosowane do możliwości inherentnych rozważaniom. Ewolucja wywarła na nas ogromną presję projektową, ale zrobiła to

bardzo nierównomiernie; presja projektowa ewolucji jest filtrowana przez niezwykłą metodologię, która działa o wiele lepiej w przypadku ręcznego masowania kodu niż w przypadku refaktoryzacji architektur programów. Teraz wyobraź sobie umysł zbudowany we własnej obecności przez inteligentnych projektantów, zaczynając od prymitywnych i niezręcznych podsystemów, które mimo to tworzą kompletny supersystem. Wyobraź sobie proces rozwoju, w którym opracowanie i okazjonalne refaktoryzowanie podsystemów może zawłaszczyć dowolny stopień inteligencji, jakkolwiek mały, wykazywany przez supersystem. Rezultatem byłby zasadniczo inny podpis projektowy i nowe podejście do sztucznej inteligencji, które nazywam seed AI. Sztuczna inteligencja (ang. seed AI) to sztuczna inteligencja zaprojektowana do samorozumienia, samomodyfikacji i rekurencyjnego samodoskonalenia. Ma to implikacje zarówno dla architektur funkcjonalnych potrzebnych do osiągnięcia prymitywnej inteligencji, jak i dla późniejszego rozwoju sztucznej inteligencji, jeśli i kiedy jej holoniczne samorozumienie zacznie się poprawiać. Sztuczna inteligencja (ang. seed AI) nie jest obejściem, które unika wyzwania ogólnej inteligencji poprzez bootstrapping z nieinteligentnego rdzenia; sztuczna inteligencja (ang. seed AI) zaczyna przynosić korzyści dopiero wtedy, gdy istnieje pewien stopień dostępnej inteligencji do wykorzystania. Późniejsze konsekwencje sztucznej inteligencji (takie jak prawdziwe rekurencyjne samodoskonalenie) pojawiają się dopiero po tym, jak sztuczna inteligencja osiągnie znaczące holoniczne zrozumienie i ogólną inteligencję. Większość tego rozdziału, sekcja 2, opisuje ogólną inteligencję, która jest warunkiem wstępnym do zasiania sztucznej inteligencji; sekcja 3 zakłada pewien stopień sukcesu w konstruowaniu ogólnej inteligencji i pyta, co może się wydarzyć później. Może to wydawać się pychą, ale można się dzięki temu dowiedzieć ciekawych rzeczy, z których niektóre implikują rozważania projektowe dotyczące wcześniejszej architektury.

Zalety umysłów w ogólności

Dla programistów komputerowych na widowni może się to wydawać oszałamiającą śmiałością, jeśli ośmielę się przewidzieć jakiekolwiek zalety SI przed ich skonstruowaniem, biorąc pod uwagę wcześniejsze niepowodzenia. Psychologowie ewolucyjni będą mniej zachwyceni, wiedząc, że pod wieloma względami ludzki umysł jest zadziwiająco wątłym dziełem. Jeśli dyskusja na temat potencjalnych zalet „SI" wydaje Ci się zbyt zuchwała, rozważ to, co następuje, nie jako dyskusję na temat potencjalnych zalet „SI", ale jako dyskusję na temat potencjalnych zalet umysłów w ogólności w porównaniu z ludźmi. Można wtedy rozważyć oddzielnie śmiałość związaną z twierdzeniem, że dane podejście do SI może osiągnąć jedną z tych zalet lub że można to zrobić w mniej niż pięćdziesiąt lat. Ludzie z pewnością posiadają następujące zalety w porównaniu z obecnymi SI:

• Jesteśmy inteligentnymi, elastycznymi, ogólnie inteligentnymi organizmami z ogromną bazą ewolucyjnej złożoności, latami doświadczenia w świecie rzeczywistym i 10^{14} równoległymi synapsami, a obecne sztuczne inteligencje takie nie są. Ludzie prawdopodobnie posiadają następujące zalety w porównaniu z inteligencjami rozwiniętymi przez człowieka dzięki przewidywalnym rozszerzeniom obecnego sprzętu:

• Biorąc pod uwagę każdy sygnał synaptyczny jako mniej więcej równoważny operacji zmiennoprzecinkowej, surowa moc obliczeniowa człowieka jest ogromnie większa niż jakiegokolwiek obecnego superkomputera lub klastrowego systemu obliczeniowego, chociaż prawo Moore'a nadal pochłania ten obszar.

• Ludzki sprzęt neuronowy – warstwa wetware – oferuje wbudowane wsparcie dla operacji takich jak rozpoznawanie wzorców, uzupełnianie wzorców, optymalizacja pod kątem powtarzających się problemów itd.; wsparcie to zostało dodane od dołu, wykorzystując mikrobiologiczne cechy neuronów,

i mogłoby być ogromnie kosztowne w symulacji obliczeniowej do tego samego stopnia wszechobecności.

• W odniesieniu do holonicznie prostszych poziomów systemu, całkowita ilość „presji projektowej" wywieranej przez ewolucję w czasie jest prawdopodobnie znacznie większa niż presja projektowa, której rozsądny zespół programistów mógłby się spodziewać osobiście.

• Ludzie mają długą historię jako inteligencje; jesteśmy sprawdzonym oprogramowaniem.

Obecne programy komputerowe niewątpliwie posiadają następujące wzajemnie synergistyczne zalety w porównaniu z ludźmi:

• Programy komputerowe mogą wykonywać bardzo powtarzalne zadania bez nudy.

• Programy komputerowe mogą wykonywać złożone, rozszerzone zadania bez popełniania tej klasy błędów ludzkich spowodowanych rozproszeniem uwagi lub przepełnieniem pamięci krótkotrwałej podczas abstrakcyjnych rozważań.

• Sprzęt komputerowy może wykonywać rozszerzone sekwencje prostych kroków z dużo większą prędkością szeregową niż ludzkie abstrakcyjne rozważania lub nawet ludzkie neurony 200 Hz.

• Programy komputerowe są w pełni konfigurowalne przez ogólne inteligencje zwane ludźmi. (Ewolucja, projektant ludzi, nie może powoływać się na ogólną inteligencję.)

Te zalety niekoniecznie przeniosą się na prawdziwą SI. Prawdziwa SI nie jest programem komputerowym bardziej niż człowiek komórką. Odpowiednia złożoność istnieje na znacznie wyższym poziomie organizacji i byłoby niewłaściwe uogólnianie stereotypowych cech komputerów na prawdziwe SI, tak jak nieodpowiednie byłoby uogólnianie stereotypowych cech ameby na współczesnych ludzi. Można powiedzieć, że prawdziwa SI zużywa moc obliczeniową, ale nie jest komputerem. To podstawowe rozróżnienie zostało zatarte przez wiele przypadków, w których etykieta „SI" została zastosowana do konstrukcji, które okazały się być jedynie programami komputerowymi; ale nadal powinniśmy oczekiwać, że rozróżnienie to będzie prawdziwe w przypadku prawdziwej SI, kiedy i jeśli zostanie osiągnięte. Potencjalne korzyści poznawcze umysłów w ogóle, w porównaniu z ludźmi, prawdopodobnie obejmują:

Nowe modalności sensoryczne: Ludzcy programiści, którym brakuje modalności sensorycznej dla języka asemblera, utknęli z abstrakcyjnym rozumowaniem i kompilatorami. Nie jesteśmy całkowicie bezradni, nawet tak daleko poza naszym środowiskiem przodków – ale tradycyjna kruchość programów komputerowych świadczy o naszej niezręczności. Umysły-w-ogóle mogą być w stanie przewyższyć ludzkie zdolności programistyczne przy stosunkowo prymitywnej ogólnej inteligencji, biorąc pod uwagę modalność sensoryczną dla kodu.

Mieszanie procesów rozmyślnych i automatycznych: Ludzkie oprogramowanie ma bardzo słabe wsparcie dla przekierowania mocy przetwarzania w czasie rzeczywistym z jednego podsystemu do drugiego. Ponadto komputer może zużywać prędkość szeregową, aby generować moc równoległą, ale neurony nie mogą zrobić odwrotnie. Umysły-w-ogóle mogą być w stanie przeprowadzić nieskomplikowany, stosunkowo mało kreatywny tok przemyślanego myślenia, używając uproszczonych procesów umysłowych, które działają z większą prędkością – idiom, który zaciera granicę między poznaniem „rozmyślnym" a „algorytmicznym". Innym przykładem zacierania się granicy jest włączanie rozważań do procesów algorytmicznych u ludzi; na przykład umysły w ogólności mogą zdecydować się na wykorzystanie inteligencji najwyższego poziomu w tworzeniu i kodowaniu jąder pojęć kategorii. Wreszcie wystarczająco inteligentna sztuczna inteligencja mogłaby włączyć de

novo funkcje programowe do procesów rozważań – tak jakby Gary Kasparow35 mógł połączyć swój mózg z komputerem i pisać drzewa wyszukiwania, aby przyczynić się do jego intuicyjnego postrzegania szachownicy.

Lepsze wsparcie dla introspektywnej percepcji i manipulacji: stosunkowo słabe wsparcie ludzkiej architektury dla introspekcji niskiego poziomu jest najbardziej widoczne w skrajnym przypadku modyfikowania kodu; możemy myśleć o myślach, ale nie o myślach o poszczególnych neuronach. Jednak inne introspekcje międzypoziomowe są również dla nas zamknięte. Brakuje nam zdolności do introspekcji jąder pojęć, alokacji skupienia uwagi, sekwencji w procesie myślowym, kształtowania pamięci, wzmacniania umiejętności itd.; nie mamy zdolności do introspekcyjnego zauważania, wywoływania przekonań na temat lub podejmowania świadomych działań w tych domenach.

Zdolność do dodawania i wchłaniania nowego sprzętu: Ludzki mózg jest ucieleśniony z typowym dla gatunku górnym limitem mocy obliczeniowej i traci neurony w miarę starzenia się. W branży komputerowej moc obliczeniowa stale staje się wykładniczo tańsza, a prędkości szeregowe wykładniczo szybsze, z wystarczającą regularnością, że „prawo Moore'a" ma rzekomo regulować jego postęp. Projekt AI nie ogranicza się również do czekania na prawo Moore'a; projekt AI, który wyświetla ważny wynik, może potencjalnie otrzymać nowe finansowanie, które umożliwi projektowi zakup znacznie większego systemu klastrowego (lub wynajęcie większej siatki obliczeniowej), być może umożliwiając AI wchłanianie setek razy większej mocy obliczeniowej. Dla porównania, 5-milionowa transformacja od Australopithecus do Homo sapiens sapiens obejmowała potrojenie pojemności czaszki w stosunku do wielkości ciała i dalsze podwojenie objętości kory przedczołowej w stosunku do oczekiwanej objętości kory przedczołowej u naczelnych z mózgiem naszej wielkości, co łącznie dało sześciokrotny wzrost pojemności kory przedczołowej w stosunku do naczelnych . Przy 18 miesiącach na podwojenie potrzeba 3,9 roku, aby prawo Moore'a objęło tak duży obszar. Nawet zakładając, że inteligencja jest bardziej oprogramowaniem niż sprzętem, to i tak jest imponujące.

Aglomeratywność: Zaawansowana sztuczna inteligencja prawdopodobnie będzie w stanie komunikować się z innymi sztuczkami z dużo większą przepustowością niż ludzie komunikują się z innymi ludźmi – w tym dzielenie się myślami, wspomnieniami i umiejętnościami w ich podstawowych reprezentacjach poznawczych. Zaawansowana sztuczna inteligencja może również zdecydować się na wewnętrzne wykorzystanie wielowątkowych procesów myślowych w celu symulacji różnych punktów widzenia. Tradycyjne, twarde rozróżnienie między „grupami" a „jednostkami" może być szczególnym przypadkiem ludzkiego poznania, a nie właściwością umysłów w ogóle. Możliwe jest nawet, że żaden projekt nigdy nie zdecydowałby się na podział dostępnego sprzętu między więcej niż jedną sztuczną inteligencję. Wiele mówi się o korzyściach płynących ze współpracy między ludźmi, ale wynika to z faktu, że istnieje gatunkowy limit na indywidualną moc mózgu. Rozwiązujemy trudne problemy, używając wielu ludzi, ponieważ nie możemy rozwiązać trudnych problemów, używając jednego dużego człowieka. Sześciu ludzi ma znaczną przewagę nad jednym człowiekiem, ale jeden człowiek ma ogromną przewagę nad sześcioma szympansami.

Sprzęt, który ma różne, ale nadal potężne zalety: Obecnym systemom obliczeniowym brakuje dobrego wbudowanego wsparcia dla biologicznych funkcji neuronowych, takich jak automatyczna optymalizacja, uzupełnianie wzorców, masowy paralelizm itp. Jednak dolna warstwa systemu komputerowego jest dobrze przystosowana do operacji takich jak refleksyjność, ślady wykonania, bezstratna serializacja, bezstratne transformacje wzorców, bardzo precyzyjne obliczenia ilościowe i algorytmy, które obejmują iterację, rekurencję i rozszerzone złożone rozgałęzienia. Również w tej kategorii, ale wystarczająco ważnej, aby zasłużyć na własną sekcję, znajduje się:

Masowy serializm: Inna „prędkość graniczna” dla prostych procesów poznawczych. Bez względu na to, jak prosty lub niedrogi obliczeniowo, prędkość ludzkiego procesu poznawczego jest ograniczona przez 200-hercową prędkość graniczną pociągów impulsów w podstawowych neuronach. Nowoczesne układy komputerowe mogą wykonywać miliardy kolejnych kroków na sekundę. Nawet jeśli SI musi „spalić” tę prędkość szeregową, aby naśladować paralelizm, proste (rutynowe, niekreatywne, nierównoległe) rozważania mogą być przeprowadzane znacznie (rzędy wielkości) szybciej niż bardziej intensywne obliczeniowo procesy myślowe. Jeśli SI ma wystarczająco dużo sprzętu lub jeśli SI jest wystarczająco zoptymalizowana, możliwe jest, że nawet pełna inteligencja SI może działać znacznie szybciej niż ludzkie rozważania.

Wolność od ewolucyjnych błędnych optymalizacji: Termin „błędna optymalizacja” tutaj wskazuje na ewolucyjną cechę, która była adaptacyjna dla inkluzywnej sprawności reprodukcyjnej w środowisku przodków, ale która dzisiaj jest w konflikcie z celami wyznawanymi przez współczesnych ludzi. Gdybyśmy mogli modyfikować nasz własny kod źródłowy, jedlibyśmy batony sałatkowe Hershey's, cieszylibyśmy się pobytem na bieżni i używalibyśmy kontroli głośności na „nudę” w czasie rozliczeń podatkowych. Wszystko, o czym ewolucja po prostu nie pomyślała: Ta ogólna kategoria jest drugą stroną ludzkiej przewagi „przetestowanego oprogramowania” – ludzie niekoniecznie są dobrym oprogramowaniem, po prostu starym oprogramowaniem. Ewolucja nie może tworzyć ulepszeń projektu, które przezwyciężają jednoczesne zależności, chyba że istnieje ścieżka przyrostowa, a nawet wtedy nie wykona tych ulepszeń projektu, chyba że ta konkretna ścieżka przyrostowa okaże się adaptacyjna z innych powodów. Ewolucja nie wykazuje żadnej przewidywalnej dalekowzroczności i jest silnie ograniczona przez potrzebę zachowania istniejącej złożoności. Ludzcy programiści mogą być kreatywni.

Rekurencyjne samodoskonalenie: Jeśli sztuczna inteligencja może się udoskonalić, każda lokalna poprawa cechy projektu oznacza, że sztuczna inteligencja jest teraz częściowo źródłem tej cechy, we współpracy z pierwotnymi programistami. Ulepszenia sztucznej inteligencji są teraz ulepszeniami źródła cechy i mogą zatem wywołać dalsze ulepszenia tej cechy. Podobnie, gdy idiom sztucznej inteligencji oznacza, że talent poznawczy przejmuje kompetencję domenową w wewnętrznych manipulacjach, ulepszenia inteligencji mogą poprawić kompetencję domenową, a tym samym poprawić talent poznawczy. Z szerszej perspektywy, samodoskonalenie umysłu jako całości może skutkować wyższym poziomem inteligencji, a tym samym zwiększoną zdolnością do inicjowania nowych samodoskonaleń.

Rekurencyjne samodoskonalenie

Całkowicie rekurencyjne samodoskonalenie jest potencjalną zaletą umysłów w ogóle, która nie ma odpowiednika w naturze – nie tylko nie ma odpowiednika w ludzkiej inteligencji, ale nie ma odpowiednika w żadnym znanym procesie. Od czasu rozbieżności rodziny człowiekowatych w obrębie rzędu naczelnych, dalszy rozwój następował w przyspieszonym tempie – ale nie dlatego, że charakter procesu ewolucyjnego uległ zmianie lub stał się „mądrzejszy”; kolejne adaptacje inteligencji i języka otworzyły nowe możliwości projektowania, a także miały tendencję do zwiększania presji selekcyjnej na inteligencję i język. Podobnie, wykładniczo przyspieszający wzrost wiedzy kulturowej u Homo sapiens sapiens został wywołany przez podstawową zmianę w ludzkim mózgu, ale sam nie miał czasu, aby wywołać jakiekolwiek znaczące zmiany w ludzkim mózgu. Gdy pojawił się Homo sapiens sapiens, późniejsze niekontrolowane przyspieszenie wiedzy kulturowej miało miejsce przy zasadniczo stałym oprogramowaniu mózgowym. Wykładniczy wzrost kultury występuje, ponieważ zdobywanie nowej wiedzy ułatwia zdobywanie większej wiedzy. Przyspieszający rozwój rodziny hominidów i wykładniczy wzrost kultury ludzkiej to dwa przykłady słabo samodoskonalących się procesów, charakteryzujących się zewnętrznie stałym procesem (ewolucja, współczesne mózgi ludzkie) działającym na pulę

złożoności (geny hominidów, wiedza kulturowa), której elementy oddziałują synergicznie. Jeśli podzielimy proces na ulepszającego i bazę treści, wówczas słabo samodoskonalące się procesy charakteryzują się zewnętrznym procesem ulepszania z mniej więcej stałą charakterystyczną inteligencją i bazą treści, w której zachodzi pozytywne sprzężenie zwrotne pod dynamiką narzuconą przez proces zewnętrzny. Jeśli sztuczna inteligencja nasion zacznie się poprawiać, będzie to oznaczać początek samooczyszczania się sztucznej inteligencji. Jakikolwiek składnik, który ulepszy sztuczna inteligencja, nie będzie już powodowany wyłącznie przez ludzi; przyczyną tego składnika stanie się, przynajmniej częściowo, sztuczna inteligencja. Każda poprawa sztucznej inteligencji będzie ulepszeniem przyczyny składnika sztucznej inteligencji. Jeśli AI zostanie ulepszona dalej – albo przez zewnętrznych programistów, albo przez wewnętrzne samodoskonalenie – AI może mieć szansę na ponowne udoskonalenie tego komponentu. Oznacza to, że wszelkie udoskonalenia globalnej inteligencji AI mogą pośrednio skutkować udoskonaleniem lokalnych komponentów przez AI. To wtórne udoskonalenie niekoniecznie umożliwia AI przeprowadzenie kolejnej, trzeciorzędnej rundy udoskonaleń. Jeśli tylko kilka małych komponentów zostało samouszczelnionych, wówczas wtórne efekty samodoskonalenia prawdopodobnie będą niewielkie, nie tego samego rzędu, co udoskonalenia wprowadzone przez ludzkich programistów. Jeśli podsystemy obliczeniowe dają początek talentom poznawczym, a talenty poznawcze plus nabyta wiedza specjalistyczna dają początek kompetencjom domenowym, to samodoskonalenie jest środkiem, dzięki któremu kompetencje domenowe mogą otaczać i ulepszać podsystemy obliczeniowe, tak jak idiom seed AI polegający na włączaniu funkcji deliberatywnych do poznania umożliwia usprawnienie kompetencji domenowych, aby otaczać i ulepszać talenty poznawcze, a zwykły idiom inteligentnego uczenia się umożliwia kompetencjom domenowym, aby otaczać i ulepszać nabytą wiedzę specjalistyczną36. Stopień, w jakim kompetencje domenowe ulepszają podstawowe procesy, będzie zależał od stopnia zaawansowania AI; stopniowo bardziej zaawansowana inteligencja jest wymagana do ulepszania wiedzy specjalistycznej, talentów poznawczych i podsystemów obliczeniowych. Stopień, w jakim poprawa inteligencji kaskadowo prowadzi do dalszych ulepszeń, będzie określany przez to, ile samouszczelnienia nastąpiło już na różnych poziomach systemu. Seed AI to silnie samodoskonalący się proces, charakteryzujący się ulepszeniami bazy treści, które wywierają bezpośredni pozytywny sprzężenie zwrotne na inteligencję podstawowego procesu ulepszania. Wykładniczy wzrost ludzkiej wiedzy kulturowej był napędzany działaniem już potężnej, ale stałej siły, ludzkiej inteligencji, na synergistyczną bazę treści wiedzy kulturowej. Ponieważ silne samodoskonalenie w sztucznej inteligencji początkowej obejmuje początkowo bardzo słabą, ale poprawiającą się inteligencję, nie można wnioskować z analogii do ludzkiego postępu kulturowego, że silnie rekurencyjne samodoskonalenie będzie podlegać wykładniczej dolnej granicy na wczesnych etapach, ani że będzie podlegać wykładniczej górnej granicy na późniejszych etapach. Silne samodoskonalenie jest mieszanym błogosławieństwem w rozwoju. We wcześniejszych epokach sztucznej inteligencji początkowej podwójny proces doskonalenia programistów i samodoskonalenia prawdopodobnie sumuje się do procesu całkowicie zdominowanego przez ludzkich programistów. Nie możemy polegać na wykładniczym bootstrappingu z nieinteligentnego rdzenia. Jednak możemy być w stanie osiągnąć potężne wyniki, bootstrappingując z inteligentnego rdzenia, jeśli i kiedy taki rdzeń zostanie osiągnięty. Rekurencyjne samodoskonalenie jest konsekwencją sztucznej inteligencji początkowej, a nie tanim sposobem na osiągnięcie sztucznej inteligencji. Możliwe, że samodoskonalenie stanie się poznawczo istotne stosunkowo wcześnie w rozwoju, ale kompleks kompetencji domenowych w celu poprawy wiedzy specjalistycznej, poznania i podsystemów nie oznacza silnych efektów rekurencyjnego samodoskonalenia. Precyzja w omawianiu początkowych trajektorii AI wymaga rozróżnienia między epokami holonicznego rozumienia, epokami rozwoju zdominowanego przez programistów i zdominowanego przez AI, epokami rekurencyjnego i nierekurencyjnego samodoskonalenia oraz epokami ogólnej inteligencji. (Czytelnicy uczuleni na zaawansowaną dyskusję na temat zaawansowanej AI mogą uznać te epoki za odnoszące się do

umysłów w ogólności, które posiadają fizyczny dostęp do własnego kodu i pewien stopień ogólnej inteligencji, za pomocą której mogą nim manipulować; uzasadnienie rozróżniania epok można rozpatrywać oddzielnie od zuchwałości sugerowania, że AI może przejść do dowolnej epoki). Epoki holonicznego rozumienia i holonicznego programowania:

1. Pierwsza epoka: AI może przekształcać kod w sposób, który nie wpływa na zaimplementowany algorytm. („Rozumienie" na poziomie kompilatora optymalizującego; tj. nie „rozumienie" w żadnym rzeczywistym sensie.)

2. Druga epoka: AI może przekształcać algorytmy w sposób, który pasuje do prostych abstrakcyjnych przekonań na temat celów projektowania kodu. Innymi słowy, AI rozumiałaby, co mają wspólnego stos zaimplementowany jako lista powiązana i stos zaimplementowany jako tablica. (Należy zauważyć, że jest to już poza zasięgiem obecnej AI, przynajmniej jeśli chcesz, aby AI sama to rozgryzła.)

3. Trzecia epoka: AI może narysować linię holoniczną od prostych wewnętrznych metryk użyteczności poznawczej (jak szybko koncepcja jest wskazywana, użyteczność koncepcji zwrócona) do określonych algorytmów. W związku z tym AI miałaby teoretyczną zdolność do wymyślania i testowania nowych algorytmów. Nie oznacza to, że AI miałaby zdolność do wymyślania dobrych algorytmów lub lepszych algorytmów, po prostu, że wynalazek w tej dziedzinie byłby teoretycznie możliwy. (Teoretyczna zdolność SI do wynalazczości nie oznacza zdolności do udoskonalania ponad wysiłki programistów. Jest to określane przez względne kompetencje domenowe i względny wysiłek włożony w dany punkt centralny.)

4. Czwarta epoka: SI ma koncepcję „inteligencji" jako produktu końcowego ciągłego holonicznego supersystemu. SI może narysować ciągłą linię od (a) swojego abstrakcyjnego rozumienia inteligencji do (b) swojego introspekcyjnego rozumienia poznania do (c) swojego rozumienia kodu źródłowego i przechowywanych danych. SI byłaby w stanie wymyślić algorytm lub proces poznawczy, który przyczynia się do inteligencji w nowy sposób i zintegrować ten proces z systemem. (Ponownie, nie oznacza to automatycznie, że wynalazki SI są ulepszeniami w stosunku do istniejących procesów.)

Epoki rzadkiego, ciągłego i rekurencyjnego samodoskonalenia:

1. Pierwsza epoka: AI ma ograniczony zestaw sztywnych procedur, które stosuje jednolicie. Gdy wszystkie widoczne możliwości zostaną wyczerpane, procedury zostają wykorzystane. Jest to zasadniczo analogiczne do zewnętrznie napędzanej poprawy kompilatora optymalizującego. Kompilator optymalizujący może wprowadzić dużą liczbę ulepszeń, ale nie są to samodoskonalenia ani ulepszenia projektowe. Kompilator optymalizujący modyfikuje język asemblera, ale pozostawia program niezmiennym.

2. Druga epoka: Procesy poznawcze, które tworzą ulepszenia, mają charakterystyczną złożoność rzędu klasycznego drzewa wyszukiwania, a nie rzędu kompilatora optymalizującego. Wystarczające inwestycje w moc obliczeniową mogą czasami przynieść dodatkowe ulepszenia, ale jest to zasadniczo wykładnicza inwestycja w liniową poprawę i bez względu na to, ile mocy obliczeniowej zostanie zainwestowane, całkowity rodzaj możliwych ulepszeń jest ograniczony.

3. Trzecia epoka: Złożoność poznawcza w domenie kompetencji AI do programowania jest na tyle wysoka, że w dowolnym momencie istnieje duża liczba widocznych możliwości złożonych ulepszeń, choć być może drobnych ulepszeń. AI zazwyczaj nie wyczerpuje wszystkich widocznych możliwości, zanim programiści nie ulepszą jej na tyle, aby nowe ulepszenia stały się widoczne. Jednak tylko ulepszenia w zakresie inteligencji napędzane przez programistów są wystarczająco silne, aby otworzyć nowe wolumeny przestrzeni projektowej.

4. Czwarta epoka: Samodoskonalenie czasami skutkuje prawdziwymi ulepszeniami w zakresie „inteligencji", „kreatywności" lub „holonicznego zrozumienia", wystarczającymi, aby otworzyć nowy wolumen przestrzeni projektowej i uczynić widoczne nowe możliwe ulepszenia.

Epoki względnej poprawy napędzanej przez człowieka i przez sztuczną inteligencję

1. Pierwsza epoka: Sztuczna inteligencja może dokonywać optymalizacji co najwyżej rzędu kompilatora optymalizującego i nie może dokonywać ulepszeń projektu ani zwiększać złożoności funkcjonalnej. Połączenie sztucznej inteligencji i programisty nie jest zauważalnie bardziej efektywne niż programista uzbrojony w zwykły kompilator optymalizujący.

2. Druga epoka: Sztuczna inteligencja może zrozumieć niewielką liczbę komponentów i wprowadzać w nich ulepszenia, ale całkowita ilość ulepszeń napędzanych przez sztuczną inteligencję jest niewielka w porównaniu z rozwojem napędzanym przez programistów. Wystarczająco duże ulepszenia programistyczne bardzo rzadko wyzwalają drugorzędne ulepszenia. Całkowita ilość pracy wykonanej przez sztuczną inteligencję nad jej własnymi podsystemami służy jedynie jako miara postępu i nie przyspiesza znacząco pracy nad programowaniem sztucznej inteligencji.

3. Trzecia epoka: Ulepszenia napędzane przez sztuczną inteligencję są znaczące, ale rozwój jest „silnie" zdominowany przez programistów w tym sensie, że ogólny postęp systemowy jest napędzany niemal wyłącznie przez kreatywność programistów. Sztuczna inteligencja mogła przejąć znaczną część pracy od programistów. Kompetencje domenowe AI w zakresie programowania mogą odgrywać kluczową rolę w dalszym funkcjonowaniu AI.

4. Czwarta epoka: Ulepszenia napędzane przez AI są znaczące, ale rozwój jest „słabo" zdominowany przez programistów. Ulepszenia napędzane przez AI i ulepszenia napędzane przez programistów są mniej więcej tego samego rodzaju, ale programiści są w tym lepsi. Alternatywnie, programiści mają więcej subiektywnego czasu na wprowadzanie ulepszeń, ze względu na liczbę programistów lub powolność AI.

Epoki dla ogólnej inteligencji

1. AI na poziomie narzędzi: zachowania AI są natychmiast i bezpośrednio określane przez programistów lub AI „uczy się" w pojedynczej domenie, korzystając z wstępnie określonych algorytmów uczenia się. (Moim zdaniem AI na poziomie narzędzi jako rzekomy krok na drodze do bardziej złożonej AI jest mocno przereklamowane.)

2. AI przedludzka: inteligencja AI nie jest znaczącym podzbiorem inteligencji ludzkiej. Niemniej jednak AI jest poznawczym supersystemem z niektórymi podsystemami, które moglibyśmy rozpoznać, i przynajmniej niektórymi zachowaniami podobnymi do umysłu. Toster nie kwalifikuje się jako „przedludzki kucharz", ale zwykły robot kuchenny mógłby to zrobić.

3. AI podludzka: inteligencja AI ma ogólnie taki sam podstawowy charakter jak inteligencja ludzka, ale jest znacznie gorsza. AI może wyróżniać się w kilku domenach, w których posiada nowe modalności sensoryczne lub inne zalety oprogramowania mózgowego, niedostępne dla ludzi. Uważam, że wartościowym testem infrahumanizmu jest to, czy ludzie rozmawiający z AI rozpoznają umysł po drugiej stronie. (Sztuczna inteligencja, która nie posiada nawet podstawowej zdolności komunikowania się z umysłami zewnętrznymi i modelowania ich, a także nie można jej nauczyć, aby to robiła, nie kwalifikuje się jako infrahuman.)

Należy ponownie podkreślić, że cała ta dyskusja zakłada, że problem zbudowania ogólnej inteligencji jest rozwiązywalny. Bez znaczącej istniejącej inteligencji domniemana „AI" pozostanie na stałe

uwięziona w pierwszej epoce programowania holonicznego – pozostanie niczym więcej niż optymalizującym kompilatorem. To prawda, że jak dotąd próby stworzenia inteligencji opartej na komputerach zawiodły i być może istnieje bariera, która stanowi, że podczas gdy 750 megabajtów DNA może określić systemy fizyczne, które uczą się, rozumują i wykazują ogólną inteligencję, żadna ilość ludzkiego projektu nie jest w stanie zrobić tego samego. Ale jeśli taka bariera nie istnieje – jeśli możliwe jest, aby sztuczny system pasował do DNA i wykazywał ogólną inteligencję równoważną człowiekowi – to wydaje się bardzo prawdopodobne, że sztuczna inteligencja zalążkowa jest również osiągalna. Byłoby szczytem biologicznego szowinizmu stwierdzenie, że podczas gdy ludzie mogą zbudować AI i ulepszyć tę AI do punktu mniej więcej równoważnej ludzkiej inteligencji ogólnej, ta sama równoważna człowiekowi AI nigdy nie opanuje (rozwiązanego przez człowieka) problemu programowania polegającego na wprowadzaniu ulepszeń do kodu źródłowego AI. Ponadto powyższe stwierdzenie błędnie przedstawia prawdopodobne powiązanie epok. AI nie musi czekać na pełną równoważność człowieka, aby zacząć ulepszać pracę programisty. Kompilator optymalizujący może „ulepszyć" pracę człowieka, poświęcając większy względny wysiłek na poziomie języka asemblera. Oznacza to, że kompilator optymalizujący wykorzystuje programowe zalety większej szybkości szeregowej i odporności na nudę, aby zastosować znacznie większą presję projektową na poziomie języka asemblera, niż człowiek mógłby wywrzeć w takim samym czasie. Nawet kompilator optymalizujący może nie dorównać człowiekowi w ręcznym masowaniu małego fragmentu języka asemblera krytycznego czasowo. Jednak przynajmniej w dzisiejszych środowiskach programistycznych ludzie nie masują już ręcznie większości kodu – częściowo dlatego, że zadanie to najlepiej pozostawić kompilatorom optymalizującym, a częściowo dlatego, że jest to niezwykle nudne i nie przyniosłoby wielu korzyści w porównaniu z wprowadzaniem dalszych ulepszeń wysokiego poziomu. Wystarczająco zaawansowana SI, która wykorzystuje asystywny serializm i wolność od ewolucyjnych błędnych optymalizacji, może być w stanie zastosować ogromne naciski projektowe na wyższe holoniczne poziomy systemu. Nawet w najlepszym wydaniu ludzie nie są zbyt dobrymi programistami; programowanie nie jest zadaniem powszechnie spotykanym w środowisku przodków. Ludzki programista jest metaforycznie niewidomym malarzem – nie tylko niewidomym malarzem, ale malarzem całkowicie pozbawionym kory wzrokowej. Tworzymy nasze programy jak artysta rysujący jeden piksel na raz, a w konsekwencji nasze programy są kruche. Jeśli ludzcy programiści SI opanują podstawowy wzorzec projektowy modalności sensorycznych, mogą obdarzyć SI modalnością sensoryczną dla struktur podobnych do kodu. Taka modalność mogłaby percepcyjnie interpretować: uproszczony język interpretowany używany do nauczania podstawowych pojęć; wszelkie wewnętrzne języki proceduralne używane przez procesy poznawcze; język programowania, w którym napisany jest poziom kodu SI; i wreszcie natywny kod maszynowy sprzętu SI. AI, która wykorzystuje modalność kodową, może nie musieć czekać na ogólną inteligencję równoważną człowiekowi, aby pokonać człowieka w określonej domenie kompetencji programowania. Nieformalnie, AI jest rodzima dla świata programowania, a człowiek nie. Prowadzi to nieuchronnie do pytania, ile umiejętności programistycznych będzie wykazywała AI zalążkowa z ogólną inteligencją równoważną człowiekowi plus modalność kodową. Niestety, prowadzi to do terytorium, które jest ogólnie uważane za tabu w dziedzinie AI. Niektórzy czytelnicy mogli zauważyć widoczną niekompletność w powyższej liście epok AI zalążkowych; na przykład ostatnim etapem wymienionym dla udoskonaleń napędzanych przez człowieka i napędzanych przez AI jest „słaba dominacja" procesu udoskonalania przez programistów ludzkich (AI i programiści wprowadzają ten sam rodzaj udoskonaleń, ale programiści wprowadzają więcej udoskonaleń niż AI). Oczywistą następną epoką jest taka, w której rozwój napędzany przez AI jest mniej więcej równy rozwojowi ludzkiemu, a epoka po niej to epoka, w której rozwój napędzany przez AI przewyższa rozwój napędzany przez człowieka. Podobnie dyskusja o epokach rekurencyjnego samodoskonalenia kończy się w punkcie, w którym napędzane przez SI udoskonalanie czasami otwiera nowe części krajobrazu możliwości, ale nie omawia możliwości samodoskonalenia bez końca: punktu,

poza którym postęp może trwać bez udziału ludzkich programistów, tak aby w momencie, gdy SI zużyje wszystkie widoczne na danym poziomie udoskonalenia, udoskonalenie to wystarczyło, aby „wspiąć się na kolejny szczebel drabiny inteligencji" i uczynić widocznym nowy zestaw udoskonaleń. Epoki ogólnej inteligencji definiują poziom narzędzi, przedludzką i podludzką SI, ale nie definiują równoważności ludzkiej ani transhumanizmu.

Infrahumanity i transhumanity: „Human-Equivalence" jako antropocentryzm

Ciekawe jest porównanie odrębnych perspektyw współczesnych badaczy sztucznej inteligencji i współczesnych psychologów ewolucyjnych w odniesieniu do konkretnego poziomu inteligencji wykazywanego przez Homo sapiens sapiens. Współcześni badacze sztucznej inteligencji są bardzo niechętni do omawiania ludzkiej równoważności, a tym bardziej tego, co może leżeć poza nią, w wyniku wcześniejszych twierdzeń o „ludzkiej równoważności", które okazały się nietrafione. Nawet wśród tych nielicznych badaczy sztucznej inteligencji, którzy nadal są skłonni omawiać ogólne poznanie, podejście wydaje się być takie: „Najpierw osiągniemy ogólne poznanie, a potem będziemy mówić o ludzkiej równoważności. A co do transhumanizmu, zapomnij o tym". Z kolei współcześni teoretycy ewolucji są mocno przeszkoleni przeciwko panglossowskim lub antropocentrycznym poglądom na ewolucję, tj. tym, w których ludzkość zajmuje jakieś szczególne lub najlepsze miejsce w ewolucji. Tutaj społecznie niedopuszczalne jest sugerowanie, że Homo sapiens sapiens reprezentuje poznanie w optymalnej lub maksymalnie rozwiniętej formie; w dziedzinie psychologii ewolucyjnej, wystająca przeszłość jest optymizmem Panglossa. Zamiast modelować rząd naczelnych i rodzinę hominidów jako ewoluujące w kierunku współczesnej ludzkości, psychologowie ewolucyjni próbują modelować rodzinę hominidów jako ewoluującą gdzieś, która następnie zdecydowała się nazwać siebie „ludzkością". (Ten pogląd jest pięknie wyjaśniony w „The Symbolic Species" Terrence'a Deacona). Patrząc wstecz na historię rodziny hominidów i linii ludzkiej, nie ma powodu, aby sądzić, że ewolucja osiągnęła twardy górny limit. Homo sapiens istniał przez krótki czas w porównaniu z bezpośrednio poprzednim gatunkiem, Homo erectus. Spoglądamy wstecz na naszą historię ewolucyjną z tego punktu widzenia, nie dlatego, że ewolucja zatrzymała się w tym punkcie, ale dlatego, że podgatunek Homo sapiens sapiens jest pierwszym rozwinięciem poznania naczelnych, które przekroczyło minimalną linię, która wspiera szybki wzrost kulturowy i rozwój psychologów ewolucyjnych. Obserwujemy inteligencję na poziomie ludzkim w naszym otoczeniu, nie dlatego, że ludzka inteligencja jest optymalna lub dlatego, że stanowi ona ograniczenie rozwojowe, ale z powodu zasady antropicznej; jesteśmy pierwszą inteligencją na tyle inteligentną, aby się rozejrzeć. Gdyby istniały podstawowe ograniczenia projektowe inteligencji, byłoby zadziwiającym zbiegiem okoliczności, gdyby koncentrowały się na poziomie ludzkim. Ściśle rzecz biorąc, postawy SI i psychologii ewolucyjnej nie są nie do pogodzenia. Można by twierdzić, że osiągnięcie ogólnego poznania będzie niezwykle trudne i że stanowi to natychmiastowe wyzwanie badawcze, jednocześnie twierdząc, że po osiągnięciu SI tylko nieuzasadniony antropocentryzm przewidywałby, że SI rozwiną się do poziomu ludzkiego, a następnie zatrzymają się. Ta hybrydowa pozycja to faktyczne stanowisko, które starałem się utrzymać w całym tym artykule – na przykład oddzielając dyskusję na temat epok rozwojowych i zalet umysłów w ogóle od zuchwałego pytania, czy AI może osiągnąć daną epokę lub przewagę. Ale byłoby głupotą udawać, że ogromna trudność osiągnięcia ogólnego poznania upoważnia nas do zamiatania jego ogromnych konsekwencji pod dywan. Pomimo lodowcowej powolności AI w porównaniu z bardziej podatnymi obszarami badań, sztuczna inteligencja nadal rozwija się w nieporównywalnie szybszym tempie niż ludzka inteligencja. Człowiek może mieć miliony lub setki milionów razy większą moc przetwarzania niż komputer osobisty około 2002 roku, ale moc obliczeniowa na dolara (nadal) podwaja się co osiemnaście miesięcy, a ludzka moc mózgu nie. Wielu spekulowało, czy rozwój AI równoważnej człowiekowi, niezależnie od tego, kiedy i jak długo to nastąpi, wkrótce nastąpi rozwój AI transhumanistycznej . Gdy sztuczna inteligencja już istnieje, może rozwijać się na wiele różnych sposobów; aby sztuczna inteligencja rozwinęła się do

punktu równoważności ludzkiej, a następnie pozostała w punkcie równoważności ludzkiej przez dłuższy okres, konieczne byłoby jednoczesne zablokowanie wszystkich swobód37 dokładnie na poziomie, który akurat zajmuje Homo sapiens sapiens. To zbyt wiele zbiegów okoliczności. Ponownie, obserwujemy inteligencję Homo sapiens sapiens w naszym otoczeniu, nie dlatego, że Homo sapiens sapiens reprezentuje podstawową granicę, ale dlatego, że Homo sapiens sapiens jest pierwszym podgatunkiem człowiekowatych, który przekroczył minimalną linię umożliwiającą rozwój psychologów ewolucyjnych. Nawet gdyby tak nie było – gdybyśmy na przykład teraz spoglądali wstecz na niezwykle długi okres stagnacji Homo sapiens – nadal byłoby to nieuprawnione wnioskowanie, że podstawowe granice projektowe, które obowiązują dla ewolucji działającej na neurony, obowiązują dla programistów działających na tranzystorach. Biorąc pod uwagę różne metody projektowania i różny sprzęt, byłby to znowu zbyt duży zbieg okoliczności. To samo dotyczy podwójnie AI nasion. Zachowanie silnie samodoskonalącego się procesu (umysłu z dostępem do własnego kodu źródłowego) nie jest takie samo jak zachowanie słabo samodoskonalącego się procesu (ewolucja udoskonalająca ludzi, ludzie udoskonalający wiedzę). Pytanie drabinowe dla rekurencyjnego samodoskonalenia – czy wspinanie się po jednym szczeblu daje punkt obserwacyjny, z którego widać wystarczająco dużo okazji, aby wystarczyły do osiągnięcia następnego szczebla – oznacza, że skutki nie muszą być proporcjonalne do przyczyn. Pytanie nie brzmi, jak duży wpływ ma dana poprawa, ale raczej, jak duży wpływ ma poprawa plus dalsze wyzwalane ulepszenia i ich wyzwalane ulepszenia. Jest to dosłownie efekt domina – uniwersalna metafora małych przyczyn o nieproporcjonalnych skutkach. Nasze instynkty dotyczące zachowań systemowych mogą wystarczyć, aby dać nam intuicyjne wyczucie wyników dowolnej pojedynczej poprawy, ale w tym przypadku nie pytamy o upadek pojedynczego domina, ale raczej o to, jak domino jest ułożone. Pytamy, czy przewrócenie jednego domina prawdopodobnie spowoduje pojedynczy upadek, dwa pojedyncze upadki, garstkę przewróconych domin, czy też przewróci cały łańcuch. Jeśli mogę pozwolić sobie na przyjęcie antybiegunowości „konserwatyzmu" – tj. pytanie, jak szybko rzeczy mogłyby się wydarzyć, a nie jak późno – to muszę zauważyć, że nie mamy pojęcia, gdzie znajduje się punkt otwartego samodoskonalenia, a ponadto nie mamy pojęcia, jak szybko nastąpi postęp po osiągnięciu tego punktu. Abyśmy nie przeceniali całkowitej ilości wymaganej inteligencji, należy zauważyć, że nieumyślna ewolucja ostatecznie natknęła się na ogólną inteligencję; po prostu zajęło to bardzo dużo czasu. Nie wiemy, jak bardzo udoskonalenie przyrostowych kroków ewolucji jest wymagane, aby silnie samodoskonalący się system przewrócił kostki domina o wystarczającej wielkości, aby każda z nich uruchamiała kolejną. Obecnie uważam, że najlepszą strategią rozwoju SI jest próba uzyskania ogólnego poznania jako niezbędnego warunku wstępnego osiągnięcia efektu domina. Jednak w teorii ogólne poznanie może nie być wymagane. Ewolucja poradziła sobie bez niego. (W pewnym sensie jest to niepokojące, ponieważ chociaż teoretycznie widzę, jak możliwe byłoby rozpoczęcie od rdzenia nierozważnego, nie mogę wymyślić sposobu, aby umieścić taki nierozważny system w ludzkich ramach odniesienia moralnego).

Koncepcyjnie możliwe jest, że podstawowe ograniczenie wyklucza wszelką poprawę efektywnej inteligencji poza nasz obecny poziom, ale nie mamy dowodów na poparcie takiego ograniczenia. Trudno mi uwierzyć, że ograniczenie umysłów w ogóle na wszystkich podłożach fizycznych przypadkowo ogranicza inteligencję do dokładnego poziomu pierwszego podgatunku człowiekowatych, który ewoluował do punktu rozwoju informatyków. Uważam, że równie trudno jest przypisać granice, które ograniczają silnie samodoskonalące się procesy do charakterystycznej szybkości i zachowania słabo samodoskonalących się procesów. „Ekwiwalencja człowieka", powszechnie uważana za wielkie nieosiągalne wyzwanie dla SI, jest chimerą – w sensie bycia zarówno „mitycznym stworzeniem", jak i „niezręcznym hybrydem". Infrahuman AI i transhuman AI są prawdopodobne jako spójne, trwałe byty. Ludzko-ekwiwalentna SI nie jest. Biorąc pod uwagę ogromne różnice architektoniczne i podłoża między ludźmi a SI oraz różne oczekiwane korzyści poznawcze, nie

ma obecnie podstaw do przedstawienia SI, która osiąga antropomorficzną równowagę kompetencji domenowych. Biorąc pod uwagę różnicę między słabo rekurencyjnym samodoskonaleniem a silnie rekurencyjnym samodoskonaleniem; biorąc pod uwagę efekt drabiny i efekt domina w samodoskonaleniu; biorąc pod uwagę różne ograniczające subiektywne wskaźniki neuronów i tranzystorów; biorąc pod uwagę potencjał umysłów w ogólności do rozszerzania sprzętu; i biorąc pod uwagę, że historia ewolucji nie daje podstaw do teoretyzowania, że zakres inteligencji Homo sapiens sapiens stanowi specjalną wolną strefę lub punkt graniczny w odniesieniu do rozwoju systemów poznawczych; dlatego nie ma obecnie podstaw, aby oczekiwać, że AI spędzi dłuższy okres w zakresie ogólnej inteligencji Homo sapiens sapiens. Homo sapiens sapiens nie jest centrum wszechświata poznawczego; jesteśmy niecentralnym przypadkiem szczególnym. Zgodnie ze standardową psychologią ludową, to czy zadanie jest łatwe, trudne czy ekstremalnie trudne, nie zmienia domyślnego założenia, że ludzie podejmują się zadania, ponieważ oczekują pozytywnych konsekwencji sukcesu. Badacze AI nadal próbują przybliżyć ludzkość do osiągnięcia AI. Niezależnie od tego, jak blisko lub daleko jest ten cel, krytycy AI są uprawnieni na mocy psychologii ludowej do wnioskowania, że ci badacze uważają AI za pożądane. Krytycy AI mogą zasadnie żądać natychmiastowej obrony tego przekonania, niezależnie od tego, czy AI jest uważane za oddalone o pięć czy pięćdziesiąt lat. Chociaż temat nie jest omawiany w tym artykule, osobiście dążę do ogólnego poznania jako środka do zasiania AI, a zasianie AI jako środka do transhumanistycznej AI, ponieważ wierzę, że ludzka cywilizacja odniesie duże korzyści z przekroczenia górnych granic inteligencji, które utrzymywały się przez ostatnie pięćdziesiąt tysięcy lat, a ponadto, że szybko zmierzamy do punktu, w którym musimy przełamać obecne górne granice inteligencji, aby ludzka cywilizacja przetrwała. Nie napisałbym artykułu na temat rekurencyjnie samodoskonalących się umysłów, gdybym uważał, że rekurencyjnie samodoskonalące się umysły są z natury złe, niezależnie od tego, czy spodziewałbym się, że ich zbudowanie zajmie pięćdziesiąt lat, czy pięćdziesiąt tysięcy lat.

Wnioski

Ludzie są ciekawi, jak rzeczy się zaczęły, a zwłaszcza pochodzenia rzeczy, które uważają za ważne. Oprócz zaspokojenia takiej ciekawości, opisy pochodzenia mogą zyskać szersze teoretyczne lub praktyczne zainteresowanie, gdy wykraczają poza opowiadanie o historycznych wypadkach, aby przekazać wgląd w trwalsze siły, tendencje lub źródła, z których zjawiska będące przedmiotem zainteresowania ogólniej wynikają. Opisy ewolucyjnej adaptacji robią to, gdy wyjaśniają, w jaki sposób i dlaczego złożona adaptacja pojawiła się najpierw w czasie lub w jaki sposób i dlaczego została zachowana od tego czasu, pod względem selekcji na podstawie dziedzicznej zmienności. [...] W takich przypadkach ewolucyjne opisy pochodzenia mogą dostarczyć wiele z tego, czego wczesni greccy myśliciele szukali w arche, czyli pochodzeniu – zunifikowanego zrozumienia pierwotnego ukształtowania czegoś, źródła ciągłego istnienia i leżącej u jego podstaw zasady.

Leonard D. Katz, „Ewolucyjne pochodzenie moralności”

Na okładce książki Douglasa Hofstadtera Gödel, Escher, Bach: An Eternal Golden Braid znajdują się dwa triplety – drewniane bloki wyrzeźbione tak, że trzy ortogonalne reflektory świecące przez trójwymiarowy blok rzucają trzy różne dwuwymiarowe cienie – litery „G”, „E”, „B”. Triplet jest metaforą sposobu, w jaki głębokie, leżące u podstaw zjawisko może dać początek wielu różnym zjawiskom powierzchniowym. Jest to metafora przecinających się ograniczeń, które dają początek całości, która jest głębsza niż suma wymagań, suma mnożnikowa, a nie addytywna. Jest to metafora dotarcia do solidnego rdzenia poprzez pytanie, co rzuca cienie i w jaki sposób rdzeń może być silniejszy niż cienie ze względu na swoją solidność. (W rzeczywistości sam triplet mógłby stanowić metaforę różnych metafor rzucanych przez koncepcję tripletu.) Poszukując arche inteligencji, starałem się nie przesadzać ani nie niedoceniać jego elegancji. Centralny kształt poznania to chaotyczny obiekt 4D,

który rzuca tysiące podpól nauk kognitywnych jako cienie 3D. Korzystając z relatywnie niewielu pól, z którymi mam niewielką znajomość, starałem się dojść do centralnego kształtu, który nie jest ani bardziej, ani mniej spójny, niż oczekiwalibyśmy od ewolucji jako projektanta. Użyłem poziomów organizacji jako strukturalnego wsparcia dla teorii, ale starałem się unikać przekształcania poziomów organizacji w arystotelesowskie kaftany bezpieczeństwa – umożliwiając dyskusję na temat „wierzeń", treści poznawczych, które łączą naturę struktur pojęciowych i wyuczoną złożoność; lub dyskusję na temat „sekwencji", adaptacji oprogramowania mózgowego, których funkcja jest najlepiej rozumiana na poziomie myśli. Poziomy organizacji są widocznie brzemienne ewolucyjnością i błagają o dopasowanie do konkretnych opisów ewolucji człowieka – ale nie oznacza to, że nasza ewolucyjna historia ustanowiła formalny postęp poprzez Modalności, Koncepcje i Myśli, przy czym każdy poziom był ukończony i kompletny przed przejściem do następnego. Poziomy organizacji strukturyzują funkcjonalny rozkład inteligencji; same w sobie nie są takim rozkładem. Podobnie poziomy struktury organizacyjnej opisują ewolucję człowieka, nie będąc same w sobie opisem ewolucji. Nie powinniśmy mówić, że Myśli ewoluowały z Konceptów; powinniśmy raczej rozważyć konkretną funkcję na poziomie myśli i zapytać, które konkretne funkcje na poziomie pojęć są konieczne i preadaptacyjne dla jej ewolucji. Budując tę teorię, starałem się unikać tych psychologicznych źródeł błędów, które moim zdaniem doprowadziły do wcześniejszych niepowodzeń w SI; zazdrość o fizykę, arystotelesowskie kaftany bezpieczeństwa, magiczne analogie z ludzką inteligencją i inne, zbyt liczne, aby je wymienić. Starałem się podać pewne wyjaśnienie wcześniejszych niepowodzeń SI, nie tylko w kategoriach „To jest magiczny klucz, którego cały czas nam brakowało (ujęcie drugie)", ale w kategoriach „To jest to, na co patrzyli poprzedni badacze, gdy dokonywali nadmiernego uproszczenia, to są siły psychologiczne leżące u podstaw początkowego nadmiernego uproszczenia i jego późniejszej propagacji społecznej, a to wyjaśnia funkcjonalne konsekwencje nadmiernego uproszczenia w kategoriach konkretnych późniejszych wyników, tak jak wydawały się one ludzkiemu obserwatorowi". Albo tak chciałbym powiedzieć, ale niestety, nie miałem miejsca w tym rozdziale na tak kompletne sprawozdanie. Niemniej jednak starałem się nie tylko przedstawić sprawozdanie z niektórych przeszłych niepowodzeń AI, ale także przedstawić sprawozdanie z tego, jak kolejne niepowodzenia próbowały i nie potrafiły wyjaśnić przeszłych niepowodzeń. Omówiłem tylko kilka najbardziej znanych i najlepiej zbadanych patologii AI, takich jak „problem uziemienia symboli" i „problem zdrowego rozsądku", ale robiąc to, starałem się przedstawić sprawozdania z ich konkretnych skutków i konkretnych źródeł. Pomimo powtarzających się niepowodzeń AI, a nawet pomimo powtarzających się nieudanych prób AI, aby wydostać się spod przeszłych niepowodzeń, AI nadal nie zakopała się tak głęboko, aby żadna możliwa nowa teoria nie mogła się z niej wydostać. Jeśli wykażesz, że nowa teoria nie zawiera zestawu przyczyn porażki w poprzednich teoriach – gdzie przyczyny porażki obejmują zarówno powierzchniowe błędy naukowe, jak i ukryte błędy psychologiczne, a przyczyny te łącznie wystarczają, aby wyjaśnić zaobserwowane patologie – to nie dowodzi to, że zidentyfikowałeś wszystkie stare przyczyny porażki ani że nowa teoria odniesie sukces, ale wystarcza, aby oddzielić nowe podejście od awersyjnego wzmocnienia w poprzednich próbach. Nie mogę obiecać, że DGI odniesie sukces – ale wierzę, że nawet jeśli DGI zostanie zabite, to nie smok AI go zabije, ale nowy i inny smok. Mam nadzieję, że przynajmniej pokazałem, że jako nowe podejście, sztuczna inteligencja oparta na DGI jest na tyle inna, że warto jej spróbować. Tak jak przedstawiono ją tutaj, teoria DGI ma ogromny potencjał do ekspansji. Mówiąc mniej uprzejmie, niniejszy rozdział jest zdecydowanie za krótki. Rozdział przedstawia opisowy, a nie konstruktywny opis funkcjonalnego rozkładu inteligencji; rozdział próbuje pokazać ewolucyjność, ale nie podaje konkretnego opisu ewolucji człowiekowatych; rozdział analizuje kilka przykładów przeszłych niepowodzeń, ale nie w pełni przeformułowuje historię sztucznej inteligencji. Szczególnie żałuję, że rozdział nie podaje ilości wyjaśnień tła, które są zwykle uważane za standardowe w przypadku wyjaśnień interdyscyplinarnych. Podczas składania elementów układanki nie byłem w stanie wyjaśnić każdego z nich osobom niezaznajomionym z nią. Zostałem zmuszony do przeciwnego ekstremum.

Niejednokrotnie skompresowałem czyjeś całe dzieło życia w jednym zdaniu i bibliografii, traktując je jak element układanki, który można złożyć bez dalszych wyjaśnień. Jedyną obroną, jaką mogę przedstawić, jest to, że centralny kształt inteligencji jest ogromny. Poproszono mnie o napisanie rozdziału w książce, a nie książki samej w sobie. Gdybym próbował opisać odniesienia interdyscyplinarne w tym, co zwykle uważa się za minimalny akceptowalny poziom szczegółowości, ten rozdział zamieniłby się w encyklopedię. Lepiej być oskarżonym o to, że nie udało się w pełni zintegrować części większej układanki, niż całkowicie ją pominąć. Jeśli rozdział jest niedokończony, niech przynajmniej będzie widocznie niedokończony. To przeczy konwencji literackiej, ale pomijanie aspektów poznania jest jednym z głównych grzechów AI. W AI naprawdę lepiej jest wspomnieć i nie wyjaśnić, niż nie wspomnieć i nie wyjaśnić – i przy tym nadal byłem zmuszony pominąć pewne rzeczy. Więc wszystkim tym, których teorie zlekceważyłem, traktując je znacznie krócej, niż na to zasługują, przepraszam. Jeśli to jakaś pociecha, traktowałem swoją własną wcześniejszą pracę nie inaczej niż traktowałem twoją. Cały temat Friendly AI został pominięty – poza jednym lub dwoma przelotnymi odniesieniami do „ludzkiego układu odniesienia moralnego" – pomimo mojego odczucia, że dyskusja na temat ludzkiego układu odniesienia moralnego nie powinna być oddzielona od dyskusji na temat rekurencyjnie samodoskonalących się, ogólnie inteligentnych umysłów. Nie mogę obiecać, że powstanie książka. Na tym etapie rytualnego postępu ogólnej teorii poznania istnieją dwie możliwe ścieżki naprzód. Można przyjąć test ognia w psychologii ewolucyjnej, psychologii poznawczej i neuronauce i spróbować pokazać, że proponowane nowe wyjaśnienie jest najbardziej prawdopodobnym wyjaśnieniem wcześniej znanych dowodów i że daje przydatne nowe przewidywania. Można też przyjąć test ognia w sztucznej inteligencji i spróbować zbudować umysł. Zamierzam obrać tę drugą ścieżkę, gdy tylko moja organizacja goszcząca znajdzie fundusze, ale może to nie pozostawić wiele czasu na pisanie przyszłych prac. Mam nadzieję, że moje wysiłki w tym rozdziale posłużą do udowodnienia, że DGI jest wystarczająco obiecujące, aby było warte znacznego finansowania potrzebnego do próby kwasowej tworzenia AI, chociaż przyznaję, że moje wysiłki w tym rozdziale nie są wystarczające, aby przedstawić DGI jako mocną hipotezę w odniesieniu do środowiska akademickiego jako całości. Ten rozdział nie zostałby napisany bez wsparcia i pomocy dużej liczby osób, których nazwisk niestety nie udało mi się zgromadzić w jednym miejscu.Wszelkie drobne skazy pozostałe w tym dokumencie są oczywiście moją winą. (Wszelkie poważne, okropne błędy lub rażące błędy logiczne zostały prawdopodobnie przemycone, gdy nie patrzyłem.) Bez Singularity Institute for Artificial Intelligence ten rozdział by nie istniał. Wszystkim darczyńcom, zwolennikom i wolontariuszom Singularity Institute, moje najszczersze podziękowania, ale jeszcze z wami nie skończyliśmy. Nadal musimy zbudować sztuczną inteligencję, a żeby to się stało, potrzebujemy znacznie więcej z was. Przepraszam hordę autorów, których nieuchronnie zlekceważyłem, nie uznając ich za twórcę pomysłu lub argumentu przypadkowo powielonego w tym rozdziale; zbiór literatury z zakresu nauk kognitywnych jest zbyt obszerny, aby jedna osoba mogła osobiście znać więcej niż nieskończenie małą część. Kiedy edytowałem szkic tego rozdziału, odkryłem artykuł „Perceptual Symbol Systems" Lawrence'a Barsalou; kiedy przesyłam ten rozdział, nadal nie przeczytałem w całości artykułu Barsalou, ale przynajmniej opisuje on model, w którym koncepcje reifikują obrazy percepcyjne i wiążą się z obrazami percepcyjnymi, a w którym kombinatoryczne struktury pojęć tworzą złożone, obrazowe obrazy mentalne. Barsalou powinien otrzymać pełne uznanie za pierwszą publikację tego pomysłu, który jest jedną z głównych podstaw teoretycznych DGI. W dzisiejszym świecie powszechnie uznaje się, że mamy obowiązek omawiania kwestii moralnych i etycznych podnoszonych przez naszą pracę. Posunąłbym się o krok dalej i powiedziałbym, że mamy nie tylko obowiązek omawiania tych kwestii, ale także dochodzenia do tymczasowych odpowiedzi i kierowania naszymi działaniami w oparciu o te odpowiedzi – nadal oczekując przyszłych ulepszeń modelu etycznego, ale także chcąc podejmować działania w oparciu o najlepsze bieżące odpowiedzi. Sztuczna inteligencja jest zbyt głęboką kwestią, abyśmy nie mieli lepszej odpowiedzi na takie konkretne pytania jak „Dlaczego?" niż „Ponieważ

możemy!” lub „Muszę jakoś zarabiać na życie”. Jeśli Homo sapiens sapiens jest niecentralnym i nieoptymalnym szczególnym przypadkiem inteligencji, to świat pełen niczego innego jak Homo sapiens sapiens niekoniecznie jest najszczęśliwszym światem, w jakim moglibyśmy żyć. Przez ostatnie pięćdziesiąt tysięcy lat próbowaliśmy rozwiązywać problemy świata za pomocą inteligencji Homo sapiens sapiens. Poczyniliśmy duże postępy, ale są też problemy, które napotkaliśmy i odbiliśmy. Może czas użyć większego młotka.