Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji

Nick Bostrom

WPROWADZENIE

Celem tego artykułu jest przeprowadzenie wstępnej analizy długoterminowych implikacji strategicznych otwartości w rozwoju SI. Jak zwiększona otwartość w rozwoju SI mogłaby wpłynąć na długoterminowe konsekwencje związane z SI? Czy wartość oczekiwana tych efektów dla społeczeństwa byłaby dodatnia, czy ujemna? Ponieważ zazwyczaj niemożliwe jest udzielenie ostatecznych odpowiedzi na tego rodzaju pytania, nasze ambicje są tutaj skromniejsze: przedstawić pewne istotne uwagi i rozwinąć przemyślenia na temat ich wagi i wiarygodności. Biorąc pod uwagę niedawne zainteresowanie tematem otwartości w sztucznej inteligencji i całkowity brak (według naszej wiedzy) jakichkolwiek prac akademickich bezpośrednio zajmujących się tym zagadnieniem, wydaje się, że nawet ta skromna praca oferuje wartościowy wkład.

Otwartość w rozwoju SI może odnosić się do różnych aspektów. Na przykład moglibyśmy użyć tego wyrażenia w odniesieniu do otwartego kodu, otwartej nauki, otwartych danych lub do otwartości na temat technik bezpieczeństwa, możliwości i celów organizacyjnych lub ogólnie do niezastrzeżonego systemu rozwoju. Możliwa jest dyskusja na temat każdego z tych różnych aspektów otwartości, nie wszystkie jednak mają te same strategiczne implikacje. Jednakże, o ile nie zostało to ustalone inaczej, używany był skrót „otwartość" w odniesieniu do praktyki udostępniania w domenie publicznej, w sposób ciągły i możliwie najszybciej, całego odpowiedniego kodu źródłowego i platform oraz swobodnego publikowania na temat algorytmów oraz spostrzeżeń i pomysłów naukowych uzyskanych w trakcie badań.

Obecnie większość wiodących programistów SI działa z wysokim, jednakże nie z maksymalnym stopniem otwartości. Na przykład badacze SI z Google, Facebooka, Microsoftu i Baidu regularnie prezentują swoje najnowsze prace na konferencjach technicznych i publikują je na serwerach. Podobnie postępują naukowcy ze środowisk akademickich. Czasami, ale nie zawsze, publikacjom tym towarzyszy wydanie kodu źródłowego, co ułatwia zewnętrznym badaczom powielanie tej pracy i kontynuowanie nowej na jej podstawie. Każda z wyżej wymienionych firm opracowała i wydała na podstawie licencji open source kod źródłowy dla platform, które pomagają badaczom oraz studentom i innym zainteresowanym wdrażać architektury uczenia maszynowego. Niedawno ogłoszona inicjatywa OpenAI ma nawet jawnie wbudowaną tożsamość marki.

Polityka wielu innych firm jest bardziej tajna lub zastrzeżona, dotyczy to szczególnie tych firm, dla których sztuczna inteligencja jest bardziej zorientowana na aplikacje. Jednakże nawet najbardziej otwarta z obecnie prowadzonych prac nie jest maksymalnie otwarta. Wyższy stopień otwartości można osiągnąć, na przykład, wykorzystując stale działające kamery internetowe i mikrofony umieszczone w laboratoriach, dzięki czemu osoby z zewnątrz mogą słuchać rozmów naukowych oraz spotkań kierownictwa, a nawet aktywnie uczestniczyć w proponowaniu i omawianiu nowych pomysłów. Laboratorium może także zatrudnić konsultantów, którzy pomogą innym grupom pracującym nad podobnymi problemami. Otwartość nie jest więc zmienną binarną, ale wektorem o wielu różnych wymiarach.

WPŁYW KRÓTKO I ŚREDNIOTERMINOWY

Pomimo że niniejszy dokument koncentruje się głównie na długofalowej perspektywie, najpierw omówiono niektóre krótko – i średnioterminowe implikacje, w celu zobrazowania zróżnicowania długoterminowości. Jest to także pomocne w zrozumieniu zachowania aktorów niedbających o aspekty długoterminowe lub instrumentalnie ograniczonych względami krótko – i średnioterminowymi.

Kwestie krótkoterminowej i niezwłocznej potrzeby otwartości można z grubsza rozłożyć na dwa pytania: 1) czy otwartość prowadzi do szybszego rozwoju i wdrażania sztucznej inteligencji? oraz 2) czy pożądane jest szybsze tworzenie i wdrażanie sztucznej inteligencji? Pytania te zostały w kolejności przeanalizowane.

Czy otwartość prowadzi do szybszego rozwoju i wdrażania SI?

W perspektywie krótkoterminowej sprawa wydaje się stosunkowo prosta. Głównym krótkoterminowym efektem otwartości istniejących badań nad sztuczną inteligencją, na przykład przez otwarty kod źródłowy i umieszczanie powiązanej własności intelektualnej w domenie publicznej, byłoby przyspieszenie rozpowszechniania i zastosowania obecnych najnowocześniejszych technik. Oprogramowanie i wiedza na temat algorytmów są dobrami niekonkurencyjnymi. Ich swobodne udostępnienie umożliwiłoby korzystanie z nich większej liczbie osób przy niskich kosztach końcowych. Ze względu na obfitość informacji w domenie publicznej efekt byłby niewielki, ale jednak pozytywny.

W perspektywie średnioterminowej sprawa jest bardziej skomplikowana. Jeśli założymy, że ten średni okres będzie wystarczająco długi, aby umożliwić przeprowadzenie znaczących nowych badań i rozwinięcie ich aż do punktu praktycznego zastosowania, to konieczne jest uwzględnienie dynamicznych skutków otwartości. W szczególności trzeba wziąć pod uwagę wpływ otwartości na zachętę do inwestowania w badania i rozwój. Być może konieczne może być uwzględnienie innych skutków pośrednich, takich jak wpływ na strukturę rynku.

Na początku należy rozważyć wprowadzenie ogólnej zasady, może to być zmiana prawa własności intelektualnej, wymóg

regulacyjny lub norma kulturowa, która popychałaby twórców SI w kierunku większej otwartości. Można wtedy spodziewać się krótkoterminowych korzyści opisanych powyżej. Jednakże w myśli ekonomicznej jest również tradycja nawiązująca do twórczości Josepha Schumpetera, która wskazuje na kompromis pomiędzy wydajnością statyczną i dynamiczną. Podstawowe pomysły stanowią dobro publiczne, a przy braku (do pewnego stopnia) pozycjonowania monopolu lub siły rynkowej firma nie jest w stanie przyswoić wartości nowych pomysłów, z których się wywodzi. Z tego punktu widzenia zyski monopolistyczne, choć zmniejszają wydajność statyczną i dobrobyt w krótkim okresie, stanowią zachętę do innowacji, które mogą poprawić dynamiczną wydajność i dobrobyt w dłuższej perspektywie czasu. W związku z tym reguła, która utrudnia deweloperowi pozyskiwanie zysków monopolistycznych na podstawie generowanych pomysłów (na przykład reguła, która zniechęca do korzystania z tajemnicy handlowej lub patentów), może mieć negatywny średnioterminowy wpływ na szybkość opracowywania i wdrażania SI.

Powinniśmy zauważyć, że nie wszystkie bodźce ekonomiczne dla innowacji zniknęłyby w otwartym, niewłasnościowym systemie innowacji. Jednym z powodów, dla których firmy angażują się w otwarte niezastrzeżone badania i rozwój, jest budowanie „zdolności absorpcyjnej": prowadzenie oryginalnych badań jako sposobu budowania umiejętności i nadążania za najnowocześniejszymi rozwiązaniami. Innym powodem jest to, że kopiowanie i wdrażanie pomysłu wymaga czasu i wysiłku, więc twórca nowego pomysłu może cieszyć się okresem skutecznego monopolu, nawet jeśli pomysł jest swobodnie przekazywany, a żadna bariera prawna nie uniemożliwia jego zaadoptowania. Nawet krótki okres wyłącznego posiadania pomysłu może umożliwić jego pomysłodawcy czerpanie zysków dzięki handlowi wiedzą poufną, na przykład dzięki wiedzy o tym, że nowa technologia mająca wpływ na rynek stała się teraz możliwa. Inną zachętą do innowacji w otwartym systemie niezastrzeżonym jest to, że pomysłodawca może czerpać zyski z posiadania uzupełniających zasobów, których wartość zwiększa nowy pomysł. Na przykład firma wydobywcza, która opracowuje nową technikę wykorzystania niektórych wcześniej niedostępnych złóż rudy, może czerpać zyski ze swojego wynalazku, nawet jeśli inne firmy wydobywcze skopiują tę technikę (choć zwykle mniej, niż gdyby jej konkurenci musieli ponieść opłaty licencyjne). Podobnie firma programistyczna może oddać swoje oprogramowanie za darmo, aby zwiększyć popyt na usługi konsultingowe i wsparcie techniczne, które właśnie ta firma może w szczególności świadczyć.

Ponadto w sektorze oprogramowania typu open source znaczny wkład wnoszą osoby, które poświęcają swój wolny czas. Jednym z motywów takiego wkładu jest umożliwienie programistom wykazania swoich umiejętności, które mogą podnieść ich wartość rynkową. Taka motywacja sygnalizująca umiejętności ma – jak się wydaje – silny wpływ na wielu badaczy SI. Naukowcy wolą pracować dla organizacji, które zachęcają ich do publikowania i prezentowania swojej pracy na konferencjach technicznych, częściowo dlatego, że dzięki temu badacz buduje reputację w środowisku naukowym i wśród potencjalnych pracodawców. Motyw sygnalizujący umiejętności jest prawdopodobnie szczególnie silny wśród najzdolniejszych młodych badaczy, ponieważ mogą oni najwięcej zyskać na popisywaniu się swoimi umiejętnościami. Daje to organizacjom, które chcą zatrudnić najbardziej utalentowanych badaczy sztucznej inteligencji, powód, by zdecydować się na otwartość w sensie unikania tajemnicy handlowej, choć niekoniecznie dążenia do patentowania, co jest całkowicie niezależne od wszelkich altruistycznych obaw związanych z promowaniem postępu naukowego lub ogólnego dobrobytu.

Tak więc niektóre zachęty do innowacji ciągle istniałyby w systemie otwartości, nawet poza dotacjami publicznymi lub filantropią. Niemniej jednak możliwe jest, że inwestycje w badania i rozwój zmalałyby, gdyby wszystkie motywacje wynikające z zalet monopolu zostały usunięte z tego zestawu. Takie ograniczenie wydatków na badania i rozwój musiałoby być zrównoważone z innymi skutkami otwartości, które mogą zwiększać postęp techniczny. Na przykład system patentowy wiąże się ze znacznymi kosztami transakcyjnymi, które zostałyby wyeliminowane przy całkowicie otwartym systemie rozwoju. Innowatorzy nie musieliby wtedy przedzierać się przez „zarośla patentowe", aby wprowadzić nowy produkt na rynek. Zrzeczenie się tajemnicy handlowej i poufności ułatwiłoby przepływ informacji pomiędzy badaczami pracującymi dla różnych organizacji, redukując niepotrzebne powielanie prac i inne nieefektywności.

Biorąc pod uwagę powyższe względy równoważące, może nie być możliwe udzielenie ogólnej odpowiedzi na pytanie, czy reguła dążąca do większej otwartości pomogłaby, czy też hamowała postęp techniczny. Znak efektu będzie zależeć od kontekstu i konkretnej formy otwartości, która jest rozważana. Należy zauważyć, że nawet jeśli zwiększenie otwartości miałoby niewielki negatywny wpływ na tempo postępu, to skutki dla dobrobytu mogą być nadal pozytywne w perspektywie krótko-, a nawet średnioterminowej. Wynika to z faktu, że otwartość poprawiłaby wydajność statyczną, udostępniając produkty niewielkim kosztem, na przykład w postaci oprogramowania typu open source, oraz umożliwiając danym najnowocześniejszym możliwościom technicznym szybsze rozproszenie w całej gospodarce. Gdyby jednak nastąpił duży negatywny wpływ na tempo postępu, wówczas straty dobrobytu wynikające z tego efektu byłyby bardziej prawdopodobne niż wzrost dobrobytu wynikający ze zwiększonej wydajności statycznej, szczególnie w perspektywie dłuższego okresu.

Do tej pory rozważaliśmy skutki wprowadzenia ogólnej zasady promującej większą otwartość. Zamiast tego moglibyśmy zapytać o skutki jednostronnej decyzji jednego aktora o dążeniu do większej otwartości. Na przykład mogłoby to być laboratorium sztucznej inteligencji, które być może z powodów altruistycznych wybrało wyższy poziom otwartości, niż byłoby to optymalne z handlowego punktu widzenia. Zakładamy, że pieniądze utracone w wyniku odstąpienia od optymalnej komercyjnie polityki zostałyby wydane na konsumpcję w postaci, która nie wpływałaby na tempo postępu technologicznego. Czy taka jednostronna decyzja przyspieszyłaby postęp techniczny?

W tym przypadku można odłożyć na bok efekty motywacji, które mogłyby zmniejszyć wydatki na badania i rozwój, gdyby wzrost otwartości był wynikiem egzogenicznej zmiany norm kulturowych lub praw własności intelektualnej. Korzyści

z otwartości omówione wcześniej nadal będą się jednak pojawiać. Taki przypadek jest więc korzystniejszy dla hipotezy, że otwartość przyspiesza postęp. Można zauważyć, że środowisko akademickie, które jest mniej uzależnione od zysków monopolistycznych w porównaniu z sektorem komercyjnym, cechuje się stosunkowo silną kulturą otwartości, co socjolog Robert Merton nazwał „normą komunistyczną" i obecnie istnieje znaczący nacisk na dalsze zwiększenie otwartości.§ W takim przypadku możliwe jest skonstruowanie modeli, w których nawet jednostronna, motywowana altruistycznie decyzja dewelopera o kontynuacji otwartego rozwoju zmniejsza całkowite wydatki na badania i rozwój. Na przykład Saint-Pierre (2003) przedstawił endogeniczny model wzrostu, w którym dla niektórych wartości parametrów taka filantropijna interwencja zmniejsza tempo wzrostu i dobrobyt przez nieproporcjonalne wypieranie własnych innowacji. Nie jest więc to taka klarowna sytuacja. Podsumowując, nadal może być prawdopodobne, że filantropijny fundator badań i rozwoju przyspieszyłby postęp przez wprowadzenie otwartości nauki, przynajmniej jeśli założymy, że badania koncentrują się na kwestiach teoretycznych lub innowacjach procesowych (w przeciwieństwie do rozwoju konkretnego produktu, który bezpośrednio konkuruje z innymi komercyjnymi możliwościami).

Czy pożądany jest szybszy postęp technologiczny i wdrażanie zdolności SI?

Prowadzi to do drugiego pytania na temat krótkookresowej i niezwłocznej celowości otwartości: zakładając, że otwartość przyspieszy postęp techniczny i rozwój zdolności SI, czy byłoby to społecznie korzystne?

Oczywiste jest, że inteligencja maszyn ma duże szanse na pozytywne zastosowania w wielu sektorach gospodarki i społeczeństwa, w tym w transporcie, opiece zdrowotnej, ochronie środowiska, rozrywce, bezpieczeństwie i odkryciach naukowych. Na przykład na całym świecie szacuje się, że każdego roku w wypadkach drogowych ginie 1,2 miliona ludzi, a liczbę tę można ostatecznie obniżyć do niskiego poziomu, ponieważ pojazdy z obsługą SI mogą przejąć więcej funkcji od ludzkich kierowców.‡ Według raportu McKinseya szacuje się, że technologie związane z SI przyczynią się do ekonomicznych wpływów w wysokości kilku bilionów dolarów rocznie do 2025 roku. Pełny przegląd potencjalnych pozytywnych zastosowań jest jednak poza zakresem tej pracy.

Podobnie jak w przypadku każdej technologii ogólnego zastosowania, można zidentyfikować obawy dotyczące każdego szczególnego wykorzystania. Argumentowano na przykład, że zastosowania wojskowe SI, w tym śmiercionośna broń autonomiczna, mogą podżegać do nowych wyścigów zbrojeniowych, zwiększyć szanse na wybuch wojny lub dać terrorystom i zabójcom nowe narzędzia do szerzenia przemocy. Techniki sztucznej inteligencji można również wykorzystać do przeprowadzania cyberataków. Algorytmy rozpoznawania twarzy, analizy emocji i eksploracji danych mogą być wykorzystywane do dyskryminacji grup znajdujących się w niekorzystnej sytuacji, naruszania prywatności ludzi lub umożliwienia represyjnym reżi-mom skuteczniejszego atakowania dysydentów politycznych. Zwiększone poleganie na złożonych systemach autonomicznych w przypadku wielu podstawowych funkcji ekonomicznych i infrastrukturalnych może tworzyć nowe rodzaje ryzyka awarii systemowych lub wprowadzać nowe słabe punkty, które mogą zostać wykorzystane przez hakerów lub cyberwojowników.

O ile możliwe jest dostrajanie wyborów dotyczących otwartości w celu różnicowego przyspieszenia określonych rodzajów aplikacji sztucznej inteligencji, obawy te mogą wskazywać na potrzebę wprowadzenia wyjątków od postawy ogólnie otwartej. Na przykład otwarty kod źródłowy broni autonomicznej wydaje się niepożądany i raczej nie znane są głosy namawiające do tego. Jednakże podstawowe badania nad SI zazwyczaj nie są nastawione na takie zastosowania. Zamiast tego, jeśli tylko badania będą pomyślne, mogą dostarczyć algorytmy i techniki, które można by zastosować w bardzo szerokim zakresie aplikacji. Dotyczy to w szczególności większości prac w obecnie kluczowych obszarach, takich jak głębokie uczenie oraz uczenie wzmacniające. Praca ta jest ekscytująca właśnie dlatego, że poszukuje się ogólnych rozwiązań problemów uczenia, które występują w szerokim zakresie zadań i środowisk.

Innym często wyrażanym problemem jest to, że postęp w dziedzinie sztucznej inteligencji spowoduje dyslokacje na rynku pracy i zmniejszy szanse na zatrudnienie niektórych pracowników. Nie jest jasne, czy krótko – i średnioterminowe możliwości SI stanowią jakiekolwiek szczególne wyzwania w tym względzie, wyzwania, które nie dotyczą ogólnie automatyzacji, a nawet dużej części wszystkich zmian technologicznych, które często zmniejszają popyt na niektóre rodzaje pracy ludzkiej. Obawy o bezrobocie technologiczne nie są niczym nowym. Po rewolucji przemysłowej kraje rozwinięte przeszły z gospodarki w przeważającej mierze rolniczej na przemysłową, a później zorientowaną na usługi. Początkowa faza uprzemysło--wienia nałożyła ogromne obciążenia na znaczną część ludności.† Jednak z biegiem czasu, po wprowadzeniu nowej polityki społecznej i przedłużeniu okresu bezprecedensowego tempa wzrostu gospodarczego, industrializacja przyniosła ogromne korzyści dla dobrobytu ludzi, zyski odzwierciedlone we wskaźnikach dotyczących żywienia, zdrowia, oczekiwanej długości życia, dostępu do informacji, mobilności i innych środków. Jeśli jako przybliżenie pierwszego rzędu modelujemy wpływ postępów sztucznej inteligencji w perspektywie krótko i średnioterminowej jako kontynuację i rozszerzenie wieloletnich trendów automatyzacji i zmian technologicznych zwiększających produktywność, to oszacowalibyśmy, że wszelkie negatywne skutki dla rynku pracy zostałyby znacznie przeważone przez zyski gospodarcze. Myślenie inaczej wydaje się pociągać za sobą przyjęcie ogólnie luddystycznego stanowiska, że być może większość obecnych zmian technologicznych ma negatywny wpływ netto.

Można podobnie zauważyć w odniesieniu do powiązanej obawy, że postępy w SI mogą pogłębiać nierówności ekonomiczne. Najlepiej rozważać to w bardziej ogólnym kontekście, w ramach szerszej dyskusji na temat zmian technologicznych i nierówności. Większość współczesnych debat wokół tych kwestii przyjmuje za pewnik, że postęp technologiczny jest zasadniczo pożądany. Kontrowersje głównego nurtu ograniczają się do tego, w jaki sposób rządy i społeczeństwa powinny się dostosować w celu przyspieszenia rozwoju i szerszego rozproszenia korzyści przy jednoczesnym szczególnym zarządzaniu tymi technologiami, które mogłyby stanowić pewne wyzwanie. Warto w tym miejscu zauważyć, że otwartość w dziedzinie SI,

niezależnie od jej wpływu na szybkość rozwoju i ogólny wzrost gospodarczy, może mieć także wyraźny wpływ na nierówności. Co oczywiste, udostępnienie oprogramowania w domenie publicznej sprawia, że jest ono dostępne bezpłatnie, co może mieć pewien wyrównujący wpływ na poziom dobrobytu osiągany przez osoby w różnych segmentach dochodów, pod warunkiem, że mają niezbędny sprzęt i umiejętności do korzystania z niego i ma to związek z ich potrzebami. Oprogramowanie typu open source może również przyczynić się do zróżnicowania korzyści dla zaawansowanych technicznie użytkowników w porównaniu z oprogramowaniem komercyjnym.

Podsumowanie krótko i średnioterminowych skutków

Wiele obecnych prac w dziedzinie SI ma otwarty charakter. Wpływ różnego rodzaju jednostronnych marginalnych wzrostów otwartości na tempo postępu technicznego w dziedzinie SI jest nieco niejasny, aczkolwiek prawdopodobnie pozytywny, szczególnie w przypadku prac teoretycznych lub innowacji procesowych. Wpływ marginalnego wzrostu otwartości wywołanego przez presję egzogeniczną, taką jak zmiany norm kulturowych lub regulacji, jest niejednoznaczny na tyle, na ile udało się zbadać tę kwestię w niniejszej analizie.

Wydaje się, że krótko i średnioterminowe skutki przyspieszenia postępu w zakresie sztucznej inteligencji są zasadniczo pozytywne, głównie ze względu na rozproszone korzyści ekonomiczne w wielu sektorach. Można zidentyfikować szereg konkretnych obszarów budzących obawy, w tym zastosowania wojskowe, zastosowania do kontroli społecznej oraz ryzyko systemowe wynikające ze zwiększonego polegania na złożonych procesach autonomicznych. Jednak w przypadku wszystkich tych obaw można również przewidzieć perspektywy korzystnych skutków, które wydają się może co najmniej równie prawdopodobne. Na przykład zautomatyzowana broń może zmniejszyć szkody wobec postronnych ludzi lub zmienić czynniki geopolityczne w pozytywny sposób. Udoskonalony nadzór może powstrzymać przestępczość, terroryzm i społecznych trutni. Bardziej wyrafinowane metody analizy danych i reagowania na nie mogą pomóc w identyfikacji i ograniczeniu różnego rodzaju ryzyka systemowego. Tak więc, pomimo że obszary budzące obawy powinny zostać zgłoszone do stałego monitorowania przez decydentów, to na podstawie obecnego stanu wiedzy nie zmieniają one oceny, że szybszy rozwój sztucznej inteligencji prawdopodobnie miałby pozytywne skutki netto w perspektywie krótko – i średnioterminowej. Podobnej oceny można dokonać w odniesieniu do obaw, że rozwój SI może mieć negatywny wpływ na rynki pracy lub nierówności ekonomiczne. Niektóre korzystne skutki w tych obszarach są również prawdopodobne, a nawet jeśli byłyby zdominowane przez negatywne skutki, to zdecydowanie pozytywny wpływ szybszego wzrostu gospodarczego najprawdopodobniej przeważyłby jakikolwiek negatywny wpływ netto na te obszary. Zauważyliśmy również, że otwartość, szczególnie w postaci umieszczania technologii i oprogramowania w domenie publicznej, może mieć pozytywny wpływ na obawy związane z dystrybucją, obniżając koszty ekonomiczne dostępu użytkowników do produktów obsługujących sztuczną inteligencję. Jednakże ze względu na to, że oprogramowanie open source wypiera pewna ilość oprogramowania komercyjnego i jest bardziej dostosowane do potrzeb zaawansowanych technicznie użytkowników, nie jest to całkowicie jasne, czy wpływ na dystrybucję faworyzowałby segmenty populacji zarówno o niskich dochodach, jak i o niskich umiejętnościach.

W skrócie, jednostronne decyzje twórców sztucznej inteligencji, by stopniowo zwiększać otwartość na podstawowe badania i innowacje procesowe, prawdopodobnie miałyby pewien pozytywny wpływ społeczny w krótkim i średnim okresie, a na marginesie przyspieszyłyby postęp w dziedzinie SI. Pod innymi względami średnioterminowe strategiczne konsekwencje różnych form otwartości są jednak bardziej niejednoznaczne i niepewne, niż można by się spodziewać.

SKUTKI DŁUGOTERMINOWE

W tej części dokonano oceny długoterminowej celowość otwartości w rozwoju SI w odniesieniu do tego, w jaki sposób otwartość wpływa na następujące dwa najważniejsze problemy związane z tworzeniem niezwykle zaawansowanych (na poziomie ludzkim lub superinteligentnych) systemów sztucznej inteligencji:

- *Problem kontroli*: jak projektować systemy SI, by działały zgodnie z intencjami projektantów.
- *Problem polityczny*: jak osiągnąć sytuację, w której osoby lub instytucje wzmocnione przez taką SI wykorzystują ją w sposób promujący wspólne dobro.

Należy przeanalizować wpływ otwartości zarówno na problem kontroli, jak i na problem polityczny. W tym miejscu zidentyfikowano trzy główne ścieżki, przez które otwartość w rozwoju SI może mieć taki wpływ lub w przeciwnym razie krzyżować się z długoterminowymi względami strategicznymi: I – otwartość może przyspieszyć rozwój SI, II – otwartość może sprawić, że wyścig o rozwój SI będzie bardziej konkurencyjny, III – otwartość może promować szersze zaangażowanie.

Otwartość może przyspieszyć rozwój SI

W poprzedniej części argumentowaliśmy, że szybszy rozwój SI jest prawdopodobną konsekwencją przynajmniej niektórych form otwartości. Może to mieć strategicznie istotne skutki na kilka sposobów, jak przedstawiono to poniżej.

Wykorzystanie wcześniej zgromadzonych zalet SI

Jest to ważne, jeśli obecnie żyjący ludzie mają silnie uprzywilejowany status w stosunku do przyszłych pokoleń według własnych kryteriów decyzyjnych. Ponieważ populacja ludzka wymiera w tempie prawie 1% rocznie, nawet niewielki wpływ na datę przybycia superinteligencji może mieć istotne znaczenie decyzyjne dla takiej obiektywnej funkcji „wpływającej na osobę" (zakładając, że superinteligencja, z dużym prawdopodobieństwem, radykalnie zmniejszyłaby śmiertelność lub poprawiła dobrobyt). Wcześniejsze zainicjowanie korzyści byłoby również ważne, jeśli znacznie zredukuje się czynnik czasowy. Jednakże wcześniejsze rozpoczęcie eksploatowania korzyści nie jest wyraźnie znaczące z bezosobowo czasowoneuronowego punktu widzenia i zamiast tego wydaje się, że należy skupić się na zmniejszeniu ryzyka egzystencjalnego.

Mniej czasu na przygotowanie

Przyspieszony rozwój SI dałby światu mniej czasu na przygotowanie się do zaawansowanej SI. Może to zmniejszyć prawdopodobieństwo rozwiązania problemu kontroli. Jednym z powodów jest to, że badania dotyczące bezpieczeństwa będą najprawdopodobniej względnie otwarte, a więc nie zyskają tyle z dodatkowych przyrostów otwartości w ogólnych badaniach

nad SI, co badania nad niebezpieczną SI. Badania związane z bezpieczeństwem mogą być w ten sposób spowolnione w porównaniu z badaniami niezwiązanymi z bezpieczeństwem.

Zmniejszy to prawdopodobieństwo, że do czasu, kiedy zaawansowana SI będzie możliwa, zostanie wykonana wystarczająca ilość pracy w zakresie bezpieczeństwa. Istnieją również procesy inne niż bezpośrednie badania nad bezpieczeństwem sztucznej inteligencji, takie jak poprawa funkcji poznawczych i ulepszenie różnych metodologii, instytucji i mechanizmów koordynacji, które mogą z czasem przyczynić się do zwiększenia gotowości. Czas na wykonanie takich badań byłby znacznie zredukowany, w momencie gdyby sztuczna inteligencja została osiągalna wcześniej. Wpływ wcześniejszego rozwoju SI na problem polityczny jest trudniejszy do oszacowania, ponieważ zależy od trudnych do przewidzenia zmian w szerszym kontekście społecznym i geopolitycznym w nadchodzących dziesięcioleciach.

Zapobieganie innym zagrożeniom egzystencjalnym

Przyspieszony rozwój SI zwiększyłby szansę, że superinteligentna SI będzie w stanie zapobiec ryzyku egzystencjalnemu pochodzącemu ze źródeł innych niż SI, takich jak ryzyko, które może powstać w wyniku syntetycznej broni biologicznej, wojny nuklearnej, nanotechnologii molekularnej lub innych, jak dotąd nieprzewidzianych czynników. Taki efekt wyprzedzający zależy od pojawienia się superinteligentnej SI, która faktycznie wyeliminuje lub zmniejszy inne poważne antropogeniczne zagrożenia egzystencjalne. To, czy tak się stanie, może częściowo zależeć od tego, czy świat po osiągnięciu SI będzie wielobiegunowy czy jednobiegunowy, co zostało omówione poniżej.

Podsumowując, fakt, że otwartość może przyspieszyć rozwój SI, wydaje się pozytywny dla celów, które dają większy priorytet obecnym ludziom niż przyszłym pokoleniom i niepewnym, bezosobowym, neutralnym czasowo celom. Każdy z tych efektów wydaje się względnie słaby w porównaniu z innymi istotnymi dla strategii skutkami wynikającymi z otwartości rozwoju SI, ponieważ nie można oczekiwać, że marginalny wzrost otwartości będzie miał więcej niż niewielki wpływ na szybkość rozwoju SI.

Otwartość sprawia, żerozwój SI jest bardziej konkurencyjny

Ważną kwestią jest to, że końcowe etapy wyścigu w celu stworzenia pierwszej superinteligentnej SI prawdopodobnie będą bardziej konkurencyjne w otwartych scenariuszach rozwoju. Powodem tego jest to, że otwartość wyrównałaby niektóre zmienne, które w przeciwnym razie spowodowałyby rozproszenie poziomów zdolności lub wskaźników postępu pośród różnych badaczy SI. Jeśli wszyscy mieliby dostęp do tych samych algorytmów lub nawet tego samego kodu źródłowego, to głównym pozostałym czynnikiem, który mógłby powodować różnice w wydajności, jest nierówny dostęp do obliczeń i danych. Można się zatem spodziewać pojawienia dużej liczby aktorów, którzy będą mogli władać sztuczną inteligencją opracowaną w otwartych projektach. Zaostrzenie sytuacji konkurencyjnej może mieć następujące ważne skutki dla problemu kontroli i problemu politycznego.

Usunięcie opcji pauzy

W trudnej, konkurencyjnej sytuacji wiodący twórcy SI mogą nie być w stanie spowolnić lub zatrzymać badań bez jednoczesnej rezygnacji z przewagi nad konkurencją. Jest to szczególnie problematyczne, jeśli okaże się, że odpowiednie rozwiązanie problemu sterowania zależy od specyfiki systemu SI, do którego ma być zastosowana. Jeśli istnieje jakaś niezbędna część mechanizmu kontroli, którą można wynaleźć lub zainstalować dopiero po zaawansowanym rozwinięciu reszty systemu SI, kluczowe może być wstrzymanie postępów w uczynieniu systemu bardziej inteligentnym, do czasu zakończenia pracy dotyczących kontroli. Załóżmy na przykład, że zaprojektowanie, wdrożenie i przetestowanie rozwiązania kontrolnego wymaga sześciu miesięcy dodatkowej pracy po tym, jak reszta SI będzie w pełni funkcjonalna. Następnie, w sytuacji nasilonej konkurencji, każdy zespół, który zdecyduje się podjąć pracę dotyczącą kontroli, może po prostu zrezygnować z przewagi nad konkurentami i tym samym z możliwości wpływania na przyszłe wydarzenia na rzecz innego, mniej ostrożnego badacza. Jeśli pula potencjalnych konkurentów o niemal najnowocześniejszych możliwościach jest wystarczająco duża, można oczekiwać, że będzie ona składać się z co najmniej jednego zespołu, który byłby skłonny kontynuować rozwój superinteligentnej SI nawet bez odpowiednich zabezpieczeń. Im większa grupa konkurentów, tym trudniej byłoby im koordynować wszystkie prace, tak aby uniknąć gwałtownego wzrostu ryzyka.

Usunięcie opcji bezpieczeństwa upośledzonej wydajności

Innym przykładem, kiedy nasilona konkurencja może być problematyczna, jest mechanizm potrzebny do zapewnienia bezpieczeństwa SI zmniejszający jej efektywność. Na przykład, jeśli bezpieczna SI działa sto razy wolniej niż niebezpieczna SI lub jeśli bezpieczeństwo wymaga ograniczenia zdolności SI, to wdrożenie mechanizmów bezpieczeństwa utrudniłoby działanie. W sytuacji nasilonej konkurencji jednostronne zaakceptowanie takiego upośledzenia może oznaczać utratę przewagi. Natomiast w mniej konkurencyjnej sytuacji, takiej jak ta, w której duża koalicja ma znaczną przewagę technologiczną lub moc obliczeniową, może być wystarczająco dużo przestrzeni na to, aby lider mógł wdrożyć pewne środki bezpieczeństwa zmniejszające efektywność bez rezygnacji ze swojego przewagi na konkurentami. Poświęcenie wydajności dla bezpieczeństwa może wymagać jedynie tymczasowego zatrzymania prac, dopóki nie zostaną opracowane bardziej wyrafinowane metody kontroli, które wyeliminują wadę wydajności bezpiecznej SI. Nawet gdyby istniały nieuniknione kompromisy pomiędzy wydajnością a bezpieczeństwem lub ograniczenia etyczne uniemożliwiające niektóre rodzaje obliczeń przydatnych z punktu widzenia instrumentalnego, sytuacja byłaby możliwa do uratowania, jeśli lider miałby wystarczającą przewagę, aby móc poradzić sobie z mniej niż maksymalnie wydajną sztuczną inteligencją w pewnym okresie. W tym czasie lider mógłby osiągnąć wystarczający stopień globalnej koordynacji, na przykład przez utworzenie singletonu, tak aby trwale zapobiec uruchomieniu bardziej wydajnych, ale i mniej pożądanych form SI lub uniemożliwić takiej SI, jeśli zostanie uruchomiona, konkurowanie z bardziej pożądanymi formami SI.

Zmiejszenie prawdopodobieństwa przechwycenia przyszłości przez małą grupę

Istnieją pewne inne konsekwencje zaostrzenia konkurencji w okresie poprzedzającym stworzenie superinteligentnej SI, które są bardziej niepewne i wartościowe, ale potencjalnie

znaczące. Jedną z takich konsekwencji jest problem polityczny. Zaostrzenie sytuacji konkurencyjnej zmniejszyłoby prawdopodobieństwo, że jeden twórca SI stanie się wystarczająco silny, aby zmonopolizować korzyści płynące z zaawansowanej sztucznej inteligencji. Jest to jedna z wymienionych motywacji projektu OpenAI, wyrażona przez Elona Muska, jednego z jego założycieli:

Myślę, że najlepszą obroną przed niewłaściwym wykorzystaniem SI jest umożliwienie jak największej liczbie osób posiadania SI. Jeśli każdy miałby potencjał SI, to nie byłoby ani jednej osoby, ani niewielkiej grupy osób, które mogłyby mieć supermoce SI.

Otwartość może zatem zwiększyć prawdopodobieństwo, że preferencje wielu ludzi będą miały wpływ na przyszłość. Może to być istotna kwestia w zależności od własnych wartości i oczekiwań, na przykład dotyczących tego, jakie preferencje będą obowiązywały, gdyby przyszłość została uchwycona przez niewielką grupę.

Oddziaływanie na wpływ uprawnień status quo

Kolejną konsekwencją dla problemu politycznego jest to, że otwartość w rozwoju SI może również wpływać na to, jakiego rodzaju podmiot najprawdopodobniej osiągnie monopol (jeśli taki istnieje) lub może osiągnąć stosunkowo największy wpływ na wynik. Dostęp do mocy obliczeniowej i ewentualnie danych staje się relatywnie ważniejszy, jeśli dostęp do algorytmów lub kodu źródłowego jest wyrównany. Można oczekiwać, że ukierunkuje to wpływ na świat post-SI bardziej na bogactwo i potęgę epoki sprzed SI, ponieważ moc obliczeniowa jest dość szeroko rozpowszechniona, także na arenie międzynarodowej, zamienna z bogactwem oraz możliwa do kontrolowania przez rządy, w porównaniu z dostępem do przełomów algorytmicznych w zamkniętym scenariuszu programistycznym, który może być bardziej nierówny, stochastyczny i lokalny. Prawdopodobieństwo, że pojedyncza korporacja lub niewielka grupa osób może dokonać krytycznego przełomu algorytmicznego potrzebnego do uczynienia SI znacznie bardziej ogólną i wydajną, wydaje się większe niż prawdopodobieństwo, że pojedyncza korporacja lub niewielka grupa osób osiągnęłaby podobnie dużą przewagę przez kontrolowanie znacznej części światowej mocy obliczeniowej. Tak więc, jeśli ktoś uważa, że lepiej jest oczekiwać, że zaawansowana SI będzie kontrolowana przez istniejące rządy, elity i zwykłych ludzi, proporcjonalnie do ich istniejącego bogactwa i władzy politycznej, zamiast przez jakąś konkretną grupę, korporację lub laboratorium, która odnosi sukcesy w dziedzinie SI, to można faworyzować scenariusz, w którym sprzęt staje się głównym czynnikiem rozwoju SI. Otwartość w rozwoju SI zwiększyłaby prawdopodobieństwo takiego scenariusza.

Jednakże otwartość zmniejszyłaby również korzyści skali w laboratoriach badawczych SI, a to faworyzowałoby mniejszych twórców, którzy mogą być mniej reprezentatywni dla potencjału rozwojowego status quo. Rozważmy odwrotny przypadek: rozwój jest całkowicie zamknięty, a każdy niedoszły twórca SI musi dokonać wszystkich istotnych odkryć i zbudować wszystkie potrzebne komponenty we własnym zakresie. O ile udana architektura SI nie okaże się niezwykle prosta, reżim ten zdecydowanie sprzyjałby większym grupom rozwojowym, a szanse wygranej przez daną grupę byłyby skalowane w zależności od wielkości grupy. Natomiast jeśli rozwój byłby otwarty, a zwycięską grupą była ta, która dodałaby jedynie ostateczny wkład do całokształtu prac, to prawdopodobieństwo wygranej przez daną grupę może zamiast tego zmieniać się mniej więcej liniowo wraz z jej rozmiarem. Tak więc w scenariuszach, w których dominowałby sprzęt, a gwałtowny rozwój SI powodowany byłby przez ostateczny postęp, otwartość zwiększyłaby prawdopodobieństwo, że to mała grupa będzie zwycięska.

W konsekwencji, jeśli większe grupy badawcze, takie jak duże korporacje lub projekty krajowe, są zazwyczaj bardziej reprezentatywne lub kontrolowane przez status quo niż przypadkowo wybrana mała grupa badaczy (np. „wynalazca w garażu"), to otwartość może albo zwiększyć albo zmniejszyć wpływ potencjału status quo na wynik, w zależności od tego, czy wąskie gardło stanowi sprzęt lub oprogramowanie. Ponieważ obecnie nie jest jasne, co będzie stanowiło wąskie gardło w przyszłości, wpływ otwartości na oczekiwany stopień kontroli potencjału status quo jest niejednoznaczny.

Zmniejszenie prawdopodobieństwa wystąpienia singletonu

Singleton to światowy porządek, w którym na najwyższym poziomie organizacji istnieje jedna skoordynowana agencja decyzyjna. Innymi słowy, singleton to reżim, w ramach którego rozwiązuje się główne globalne problemy dotyczące koordynacji lub negocjacji. Pojawienie się singletonu jest zatem spójne z obydwoma scenariuszami, w których wiele ludzkich woli razem kształtuje przyszłość, oraz ze scenariuszami, w których przyszłość zostaje przejęta przez wąskie grono. Punkt, w którym otwartość w rozwoju SI wydaje się obniżać prawdopodobieństwo singletonu, jest zatem różny od punktu, w którym otwartość wydaje się zmniejszać prawdopodobieństwo opanowania przyszłości przez małą grupę. Można sprzeciwić się tej małej grupie, a jednocześnie popierać utworzenie singletonu. Istnieje szereg poważnych problemów, które mogą pojawić się w wielobiegunowym wyniku, których można by uniknąć w singeltonie.

Jednym z takich problemów jest to, iż może się okazać, że na pewnym poziomie rozwoju technologicznego i być może w dojrzałości technologicznej występek będzie miał przewagę nad obroną. Załóżmy na przykład, że wraz z dojrzewaniem biotechnologii niedrogie staje się zaprojektowanie mikroorganizmu, który może siać spustoszenie w środowisku naturalnym, którego ochrona przed uwolnieniem i namnażaniem takiego organizmu jest zbyt kosztowna. Następnie w wielobiegunowym świecie, w którym istnieje wiele niezależnych centrów inicjatywy, można by oczekiwać, że organizm zostanie ostatecznie uwolniony, być może przez przypadek, w ramach operacji szantażu, przez agenta o apokaliptycznych wartościach, albo w wyniku działań wojennych. Prawdopodobieństwo uniknięcia takiego wyniku wydaje się maleć wraz z liczbą niezależnych podmiotów, które mają dostęp do odpowiedniej biotechnologii. Ten przykład można uogólnić: nawet jeśli w biotechnologii występek nie będzie miał takiej przewagi, może tak być w cyberzbrojeniach, w nanotechnologii molekularnej, przy opracowywaniu zaawansowanych dronów bojowych, lub w jeszcze innej nieoczekiwanej technologii, która zostałaby opracowana przez superinteligentną SI. Świat, w którym problemy globalnej koordynacji pozostają nierozwiązane, nawet w momencie gdy

rozwój technologii zmierza w kierunku swoich fizycznych granic, staje się zakładnikiem możliwości, że na pewnym poziomie rozwoju technologicznego natura zbyt silnie będzie faworyzować zniszczenie nad stworzeniem. Z punktu widzenia redukcji ryzyka egzystencjalnego może być zatem korzystne wyłonienie pewnych ustaleń instytucjonalnych, które umożliwiłyby solidną globalną koordynację. Może to być łatwiejsze, jeśli początkowo mało jednostek miałoby zaawansowane zdolności rozwoju SI i potrzeby koordynacji.

Możliwość, że występek może mieć nieodłączną przewagę nad obroną, nie jest jedyną kwestią związaną z wynikiem wielobiegunowym. Innym problemem jest to, że przy braku globalnej koordynacji może nie być możliwe zapobieżenie gwałtownemu rozwojowi populacji cyfrowych umysłów i wynikającej z tego epoki maltuzjańskiej, w której może ucierpieć dobro tych cyfrowych umysłów. Niezależni aktorzy mieliby silną motywację do zwielokrotnienia liczby kontrolowanych przez nich pracowników cyfrowych do punktu, w którym krańcowy koszt produkcji kolejnego pracownika (w tym energii elektrycznej i wynajmu sprzętu) równałby się przychodom, które może uzyskać taki pracownik, pracując maksymalnie ciężko. Lokalne lub krajowe przepisy mające na celu ochronę dobrobytu umysłów cyfrowych mogą przenieść produkcję do jurysdykcji, które oferują bardziej korzystne warunki dla inwestorów. Proces ten może przebiegać szybko, ponieważ umysły cyfrowe napotykają mniej barier dla migracji niż biologiczna siła robocza, a świadczone przez niego usługi informacyjne są w dużej mierze niezależne od położenia geograficznego, choć podlegają efektom opóźnień spowodowanych transmisją sygnału na duże odległości, co może być znaczące dla umysłów cyfrowych pracujących z dużą szybkością. Długoterminowa równowaga takiego procesu jest trudna do przewidzenia i może być determinowana przede wszystkim wyborami dokonywanymi po opracowaniu zaawansowanej SI. Jednakże stworzenie stanu rzeczy, w którym świat jest zbyt podzielony i wielobiegunowy, aby móc wpływać na to, dokąd prowadzi, powinno stanowić powód do niepokoju, chyba że istnieje pewność (trudno jest dostrzec, co uzasadnia takie zaufanie), że programy z najwyższą sprawnością w dojrzałej algorytmicznej hiperekonomii są zasadniczo takie same jak programy, które mają najwyższy poziom subiektywnego dobrobytu lub wartości moralnej.

Znaczenie mnogości SI dla problemu kontroli

Można by pomyśleć, że zaostrzenie konkurencji sprzyjałoby bardziej osiągnięciu pożądanego wyniku, pomagając tym samym rozwiązać problem kontroli. Pomysł polegałby na tym, że w scenariuszu ściślej konkurencyjnym mniej prawdopodobne jest, że jeden system SI wyprzedzi wszystkie pozostałe, tak aby uzyskać decydującą przewagę strategiczną. Zamiast tego bardziej prawdopodobne byłoby istnienie wielu systemów SI zbudowanych przez różnych ludzi w różnych krajach do różnych celów, ale o porównywalnym poziomie zdolności. W takim wielobiegunowym świecie spowodowanie ekstremalnych szkód przez każdy z tych systemów SI może być trudniejsze. Byłoby tak nawet wtedy, gdyby zawiodły elementy sterujące zastosowane do tych systemów, ponieważ istniałyby inne SI, prawdopodobnie pod ludzką kontrolą, które mogłyby je powstrzymać.

Ten sposób myślenia jest dość problematyczny jako argument za otwartością, nawet jeśli odłożymy przedstawione wcześniej ogólne obawy dotyczące wielobiegunowości. Istnienie wielu SI nie gwarantuje, że będą działać w interesie ludzi lub pozostaną pod ludzką kontrolą. Można posłużyć się analogią, istnienie wielu konkurujących współczesnych ludzi w nie-wielkim stopniu przyczyniło się do tego, by promować długoterminowy dobrobyt innych gatunków hominidów, z którymi Homo sapiens kiedyś dzielił planetę. Jeśli SI byłyby kopia-mi tego samego szablonu lub jego niewielkimi modyfikacjami, wszystkie mogą zawierać tę samą wadę kontrolną. Otwarty rozwój może w rzeczywistości zwiększyć prawdopodobień-stwo takiej jednorodności, ułatwiając różnym laboratoriom korzystanie z tego samego kodu podstawowego i algorytmów zamiast wymyślania własnych.

Istnieje również możliwość awarii systemowych wynikających z nieoczekiwanych interakcji różnych SI. Wiemy, że takie

awarie mogą wystąpić nawet w przypadku bardzo prostych algorytmów (na przykład Flash Crash). Wśród zaawansowanych sztucznych agentów zdolnych do wysoce zaawansowanego planowania i strategicznego wnioskowania, które mogą być w stanie koordynować przy użyciu innych, bardziej skutecznych środków niż ludzie, mogą istnieć dodatkowe i zupełnie nowe możliwości wystąpienia awarii systemowych. Nawet jeśli jakaś równowaga sił zapobiegłaby naruszeniu ludzkich interesów przez pojedynczą SI lub koalicję SI, to nie jest jasne, czy możemy być pewni, że stan taki się utrzyma.

Gdyby, z punktu widzenia kontroli, tak naprawdę korzystne było posiadanie wielu SI, to lepszym rozwiązaniem mogłoby być utworzenie licznych SI przez jednego twórcę, który miałby większą zdolność do zapewnienia, że te liczne SI będą zrównoważone pod względem swoich zdolności. Prawdą jest jednak to, że SI stworzone przez jednego twórcę mogą być bardziej podobne do siebie, a zatem bardziej podatne na skorelowane błędy kontroli, niż SI stworzona przez różnych programistów. Jak zauważyliśmy jednak, choć otwartość może zwiększać prawdopodobieństwo istnienia wielu różnych twórców jednocześnie, to będzie ona również powodować, że SI tworzona przez nich będzie oparta na zbliżonych projektach. Z tego powodu wpływ netto otwartości na prawdopodobieństwo istnienia różnorodnego zestawu SI jest niejednoznaczny.

Moglibyśmy zebrać zestaw założeń popierających tezę, że powinno się dążyć do rozwiązania problemu kontroli przez stworzenie wielu SI w wyniku przyjęcia polityki otwartości. Na przykład moglibyśmy zastrzec, że mnogość SI, nawet jeśli byłyby one oparte na tym samym projekcie, przyczyniłaby się do bezpieczeństwa pod warunkiem określenia różnych celów dla SI. Można by wtedy argumentować, że SI stworzona przez różnych twórców w natural-ny sposób miałaby zróżnicowane cele, a tym samym przyczyniłaby się do bezpieczeństwa publicznego, podczas gdy pojedynczy twórca stworzyłby tylko jedną SI lub wiele SI o identycznych celach, ponieważ nadanie SI innego celu wiązałoby się z dodatkowym kosztem, a taka SI nie działałaby wyłącznie w interesie twórcy. Pewnym wyobrażeniem mógłby być świat, gdzie istnieje wiele SI, z których każda dążyłaby do innego celu i żadna nie byłaby wystarczająco silna, aby przejąć kontrolę sama lub tworząc koalicję z innymi silnymi SI. Takie SI konkurowałyby o klientów i inwestorów, oferując nam korzystne oferty, podobnie jak korporacje konkurujące o ludzkie względy w kapitalistycznej gospodarce.

W takim modelu należy również uwzględnić rolę państwa. Bez istnienia państwa na tyle silnego, aby regulować konkurencyjne SI oraz egzekwować prawo i porządek, wątpliwe może być to, jak długo utrzymałaby się równowaga sił i jakie miejsce w niej zajmowaliby ludzie. Alternatywną i mniej atrakcyjną analogią może być XVII-wieczna Europa, gdzie SI odpowiadałaby silniejszym państwom, a populacja ludzka odpowiadałaby niewielkim księstwom mającym nadzieję na osiągnięcie bezpieczeństwa przez połączenie się z silną i zwycięską koalicją SI.

Podsumowując, można oczekiwać, że otwartość uczyni rozwój SI bardziej konkurencyjnym, a miałoby to kilka strategicznych konsekwencji. Utrudniłoby to przerwanie prac przed ich zakończeniem w celu wdrożenia lub przetestowania mechanizmów bezpieczeństwa, jak również stosowanie jakiegokolwiek mechanizmu bezpieczeństwa, który zmniejszałby wydajność. Obie z tych konsekwencji zdają się mieć negatywne skutki dla problemu kontroli. Otwartość miałaby również konsekwencje dla problemu politycznego związane ze zmniejszeniem prawdopodobieństwa zmonopolizowania korzyści płynących z zaawansowanej SI przez małą grupę oraz wystąpienia singletonu. Mogłoby to zwiększyć lub zmniejszyć wpływ potencjału status quo na przyszłość post-SI, w zależności od tego, czy transformacja byłaby głównie ograniczona sprzętowo, czy też programowo. Co więcej, mógłby być obserwowany wpływ na problem kontroli związany z dystrybucją SI, wynikający z otwartego rozwoju, aczkolwiek wielkość i znaczenie tego wpływu są niejasne. Otwartość mogłaby zwiększyć prawdopodobieństwo mnogiej SI, co z kolei mogłoby zwiększyć prawdopodobieństwo osiągnięcia pewnego rodzaju układu równowagi sił pomiędzy SI. Z drugiej strony otwartość mogłaby również uczynić różne SI bardziej podobnymi do siebie niż w przypadku scenariusza mnogiej SI realizowanego bez otwartości, a zatem bardziej prawdopodobne byłoby wystąpienie skorelowanych niepowodzeń. W każdym razie nie jest jasne, czy mnogość różnorodnych SI stworzonych przez różnych twórców naprawdę byłaby pomocna w rozwiązaniu problemu związanego z kontrolą.

Otwartość promująca szersze zaangażowanie

Jedną z klas potencjalnie istotnych strategicznie skutków otwartości w rozwoju SI jest to, że otwartość może zwiększyć zewnętrzne zaangażowanie z różnymi aspektami najnowocześniejszej technologii SI. Taka otwartość powinna zwiększyć zainteresowanie zewnętrzne, aczkolwiek zainteresowanie nie jest aksjomatyczne. Bardzo często efektem prób zachowania czegoś w tajemnicy jest jedynie zwrócenie na to większej uwagi. Jednak w przypadkach, w których znaczące zaangażowanie wymaga szczegółowych informacji i precyzyjnego dostępu, prawdopodobne jest, że większa otwartość zwiększyłaby takie zaangażowanie.

Perspektywy zewnętrzne rzucają światło na bezpieczeństwo

Ktoś mógłby zatem argumentować, że jeśli systemy SI są utrzymywane w tajemnicy, to zewnętrzni eksperci nie mogą bezpośrednio nad nimi pracować, tak aby były bezpieczniejsze, a to w efekcie sprawia, że taki scenariusz zamkniętego rozwoju jest bardziej ryzykowny. Należy jednak pamiętać, że jeśli systemy SI są utrzymywane w tajemnicy, to zewnętrzni eksperci nie mogą również bezpośrednio pracować nad zwiększeniem ich efektywności. Na pierwszy rzut oka może to wyglądać jak remis: jeśli nie ma różnicującego wpływu na bezpieczeństwo, to sprowadza się to do tego, że otwartość może po prostu przyspieszyć badania nad bezpieczeństwem i efektywnością, które zostały opisane we wcześniejszej części. Można jednak spekulować, że praca nad bezpieczeństwem uzyskałaby więcej korzyści z udziału zewnętrznego niż praca mająca na celu zwiększenie efektywności SI. Wynikałoby to być może z tego powodu, że inżynieria bezpieczeństwa i analiza ryzyka są bardziej podatne na przemyślenia grupowe i inne uprzedzenia, a zatem skorzystałyby nieproporcjonalnie więcej w przypadku udziału zewnętrznego. Przypuszczalnie łatwiej jest łudzić się na temat bezpieczeństwa SI, którą się tworzy, niż łudzić się na temat jej zdolności, ponieważ istnieje więcej możliwości obiektywnej

informacji zwrotnej na temat tego drugiego. Z tego powodu w przypadku optymistycznego uprzedzenia istnieje większa swoboda w wypaczaniu przekonań na temat bezpieczeństwa niż efektywności SI. Dodatkowo, jeśli perspektywy zewnętrzne stanowią korektę takiego nastawienia, to ich uwzględnienie przyniesie efekt w różnicowym promowaniu postępu w zakresie bezpieczeństwa.

Uczestnicy zewnętrzni bardziej altruistyczni?

Ponadto można argumentować, że ponieważ bezpieczeństwo jest dobrem publicznym, zewnętrzni badacze oraz ich sponsorzy mają bardziej usprawnić pracę w zakresie bezpieczeństwa niż w zakresie efektywności, w odniesieniu do alokacji wysiłków wykonanych przez badaczy we własnym zakresie, którzy prawdopodobnie mają stosunkowo silniejsze niealtruistyczne motywy do pracy nad efektywnością. Otwartość w rozwoju SI mogłaby następnie przez umożliwienie udziału bezinteresownym osobom z zewnątrz zwiększyć ogólną część wysiłku związanego z rozwojem SI, który koncentruje się na bezpieczeństwie, a tym samym zwiększyć szanse, że problem kontroli zostanie na czas rozwiązany.

W przypadku grupy, która jest wystarczająco altruistyczna i zorientowana na bezpieczeń-stwo, argument ten może zostać odwrócony. W przypadku takiej grupy otwartość może osłabić koncentrację na dobrach publicznych, umożliwiając udział mniej sumiennym osobom z zewnątrz.

Wpływ na architekturę?

Możliwe, że zasady działania organizacji oparte na założeniu otwartego rozwoju mogą wpłynąć dodatnio lub negatywnie na charakter tworzonej SI. Na przykład popularne w projektach programistycznych open source podejście „rafy koralowej" mogłoby spowodować zachłanną pogoń za lokalnym optimum zamiast cierpliwego poszukiwania tego globalnego.§ Może także zaistnieć sytuacja, kiedy luźniejsze sprzężenie pomiędzy grupami twórców będzie zachęcać do większej funkcjonalnej modułowości, co w porównaniu ze scentralizowanymi procesami może sprzyjać ściślej zintegrowanej jednolitej architekturze. Możliwe jest, że takie efekty mogą mieć znaczące implikacje dla problemu kontroli, jednak niepewność, jakiego rodzaju mogą to być skutki, a także czy dany efekt byłby pozytywny, czy negatywny dla problemu kontroli, może być zbyt duża dla tego rodzaju rozważań, aby mieć znaczący wpływ na nasze obecne przemyślenia.

Udostępnienie jednostkom więcej przezorności

Otwartość na możliwości, to, do czego zdolna jest inteligencja maszynowa w danym momencie i jaki jest oczekiwany harmonogram dalszych postępów, zwiększyłoby możliwości osób z zewnątrz do wpływania lub dostosowywania się do rozwoju SI. Mogłoby to zwiększyć prawdopodobieństwo nacjonalizacji wiodących działań związanych z rozwojem SI, ponieważ ułatwiłoby to rządowi dokładne monitorowanie, kiedy i gdzie trzeba będzie interweniować w celu utrzymania kontroli nad zaawansowanymi zdolnościami SI. Z drugiej strony otwartość na naukę i kod źródłowy może zmniejszyć prawdopodobieństwo nacjonalizacji, upubliczniając rozwój SI, także na arenie międzynarodowej, i tym samym utrudniając przejęcie przez rząd. Otwartość może również zmniejszyć prawdopodobieństwo nacjonalizacji przez wspieranie kultury wśród badaczy SI, która jest bardziej nieprzychylnie nastawiona wobec rządowej lub korporacyjnej kontroli nad rozwojem SI.

Otwartość na możliwości, oprócz ułatwienia rządowej kontroli nad przełomowymi odkryciami w dziedzinie SI, byłaby również pomocna w przygotowaniu się społeczeństwa, zapewniając różnym podmiotom wyraźniejszy obraz przyszłości. Nie jest od razu jasne, jaki miałoby to wpływ na problem kontroli lub problem polityczny. Można oczekiwać, że większa przezorność dotycząca nadchodzącej rewolucji technologicznej przyniesie rozliczne pozytywne skutki, umożliwiając planowanie i adaptację. W szczególności otwartość może umożliwić dokładniejsze prognozowanie ryzyka związanego z problemem kontroli, co może prowadzić do zwiększonych inwestycji w te rozwiązania w tych krajach, w których są one szczególnie potrzebne.

Zobowiązanie do udostępniania

Omówiliśmy już, w jaki sposób otwartość uczyniłaby rozwój SI bardziej konkurencyjnym i jak mogłaby przyspieszyć postęp, a także krótkoterminowe korzyści wynikające z umożliwienia wykorzystania istniejących pomysłów i informacji niewielkim kosztem. W tym miejscu odnotowujemy kolejną możliwą strategicznie istotną konsekwencję. Otwartość w najbliższym czasie może stworzyć pewien rodzaj blokady, która zwiększy szansę, że bardziej zaawansowane możliwości SI lub przynajmniej jej niektóre elementy zostaną także powszechnie udostępnione, nawet jeśli pozostałe, np. moc obliczeniowa, pozostaną własnościowe. Taka blokada może nastąpić w przypadku zakorzenienia kulturowej normy otwartości lub jeśli poszczególni twórcy SI podejmą zobowiązania wobec otwartości, z których później nie będą mogli się łatwo wycofać. Prowadziłoby to do wspomnianych wcześniej problemów, dając obecnej otwartości możliwość zwiększania konkurencyjności rozwoju SI, a być może także w dłuższej perspektywie.

Istnieje też osobny i korzystny efekt blokady otwartości, która może sprzyjać dobrej woli i współpracy. Im bardziej różni potencjalni twórcy sztucznej inteligencji oraz ich poplecznicy czują, że w pełni podzieliliby się korzyściami sztucznej inteligencji, nawet jeśli przegrają wyścig o opracowanie SI jako pierwsi, tym mniejszą mają motywację do nadania priorytetu rozwojowi bezpieczeństwa i tym łatwiej im będzie współpracować z innymi stronami przy bezpiecznym i pokojowym rozwoju zaawansowanej SI zaprojektowanej dla dobra wspólnego. Takie podejście oparte na współpracy miałoby prawdopodobnie pozytywny wpływ zarówno na problem kontroli, jak i na problem polityczny.

Podsumowując, otwarty scenariusz rozwoju może zredukować grupowe myślenie oraz inne obciążenia w projekcie SI, umożliwiając osobom zewnętrznym większe zaangażowanie się, co może różnie wpłynąć na analizę ryzyka i inżynierię bezpieczeństwa, pomagając w ten sposób w problemach dotyczących kontroli. Udział osób postronnych może być również motywowany altruistycznie, a zatem bardziej ukierunkowany na bezpieczeństwo niż na wydajność. Otwarta współpraca może wpływać na wybory projektowe w rozwoju inteligentnych maszyn, być może sprzyjając bardziej przyrostowemu podejściu w stylu „rafy koralowej", lub zachęcać do zwiększonej modułowości, choć obecnie nie jest jasne, jak wpłynie to na problem kontroli. Otwartość na możliwości dałaby różnym podmiotom lepszy wgląd w bieżący i oczekiwany rozwój, ułatwiając planowanie i adaptację. Taka otwartość może również ułatwić wywłaszczenie przez rząd, podczas gdy otwartość dotycząca nauki i kodu przeciwdziałałaby wywłaszczeniu, pozostawiając mniej zastrzeżonych materiałów do zagarnięcia. Wreszcie, jeśli obecne wybory dotyczące otwartości podlegałyby efektom blokowania, miałyby one bezpośredni wpływ na przyszłe poziomy otwartości i mogłyby służyć jako sposoby zaangażowania się w dzielenie się efektami zaawansowanej SI, co byłoby pomocne zarówno w przypadku problemu z kontrolą, jak i problemu politycznego.

WNIOSKI I ZALECENIA

Jak zostało to omówione, strategiczne implikacje otwartości w dziedzinie SI są kwestią znacznie złożoną. Przeprowadzona analiza i wszelkie wyciągnięte z niej wnioski są zarówno niepewne, jak i wstępne, jednakże zidentyfikowano kilka istotnych rozważań, które należy wziąć pod uwagę w każdej uzasadnionej ocenie na ten temat.

Oprócz konsekwencji omówionych w tym artykule, istnieje wiele lokalnych efektów otwartości, które indywidualni twórcy SI będą chcieli wziąć pod uwagę. Projekt może czerpać prywatne korzyści z otwartości, na przykład podczas rekrutacji (badacze lubią publikować i budować reputację), umożliwiając menedżerom porównywanie wewnętrznych badań z zewnętrznymi standardami, oraz przez prezentowanie osiągnięć dla prestiżu i chwały. Efekty te nie zostały uwzględnione w niniejszej analizie, ponieważ nacisk został położony na globalną potrzebę otwartości, a nie na taktyczne zalety lub wady, jakie może ona pociągać za sobą dla określonych grup SI.

Ogólna ocena

W najbliższej perspektywie można oczekiwać otwartości na przyspieszenie rozpowszechniania istniejących technologii, co miałoby pewne ogólnie pozytywne skutki gospodarcze, a także szereg bardziej szczegółowych efektów, zarówno pozytywnych, jak i negatywnych, wynikających z konkretnych zastosowań. Jednakże oczekiwana wartość netto tych efektów byłaby dodatnia. Z perspektywy krótkoterminowej pożądana jest zatem prawie każda forma zwiększonej otwartości. Niektóre obszary zastosowania budzą szczególne obawy, w tym zastosowania wojskowe, kontroli społecznej oraz ryzyka systemowego wynikającego ze zwiększonego polegania na złożonych autonomicznych procesach. Zastosowania te powinny być omawiane przez odpowiednie zainteresowane strony i monitorowane przez decydentów w miarę gromadzenia się rzeczywistych doświadczeń dotyczących tych technologii.

Wpływ na rynki pracy można w pierwszej kolejności uwzględnić w bardziej ogólnej kategorii automatyzacji i oszczędzającego pracę postępu technologicznego, który historycznie miał ogromny pozytywny wpływ netto na dobrobyt ludzi, choć nie bez dużych kosztów przejścia dla części populacji. W przypadku znacznego wzrostu tempa lub zakresu automatyzacji można wymagać rozszerzonego wsparcia socjalnego dla wysiedleńców i innych słabszych grup społecznych. Dystrybucyjne skutki zwiększonej otwartości są nieco niejasne.

Pomimo że artykuł ten nie jest szczególnie długi, jest dość konkretny, a wiele rozważań, którym poświęcono tu tylko kilka słów, z łatwością mogłoby być przedmiotem całej osobnej analizy.

Możliwe jest również, że część struktury niniejszej analizy jest istotna dla innych zagadnień makrostrategicznych i w ten sposób mogłaby ukierunkowywać na szerszy zestaw zagadnień.

Historycznie oprogramowanie open source było szczególnie popularne wśród zaawansowanych technicznie użytkowników, jednakże mniej wykwalifikowani użytkownicy również odnieśliby korzyści na przykład z produktów zbudowanych na oprogramowaniu open source lub za pomocą zaawansowanych użytkowników jako pośredników.

Średnioterminowe skutki otwartości komplikuje możliwość wpływu na motywację do innowacji lub strukturę rynku. Literatura na temat ekonomii innowacyjnej jest tu istotna, ale niejednoznaczna. Można przypuszczać, że jednostronny wzrost otwartości ma pozytywny wpływ na tempo postępu technicznego w dziedzinie SI, zwłaszcza jeśli badania koncentrują się na pracy teoretycznej lub innowacjach procesowych.

Wpływ wzrostu otwartości spowodowanego presją egzogeniczną, na przykład wynikającą z przepisów lub norm kulturowych, jest niejednoznaczny. Średnioterminowy wpływ szybszego postępu technicznego w zakresie SI można ocenić w podobny sposób jak wpływ krótkoterminowy. Istnieją zarówno pozytywne, jak i negatywne zastosowania oraz wiele niepewności, jednak uzasadnione jest przypuszczenie, że rezultat netto skutków średniookresowych będzie pozytywny. Można tak wnioskować, wykorzystując ekstrapolację przeszłego postępu technologicznego i wzrostu gospodarczego. Potencjalne obawy w perspektywie średnioterminowej dotyczą nowej formy zaawansowanej wojny robotycznej, która może obejmować destabilizujące wydarzenia, takie jak wyzwania związane z odstraszaniem nuklearnym, na przykład od autonomicznych botów śledzących okręty podwodne lub głęboką infiltrację terytorium wroga przez małe systemy robotyczne oraz użycie sztucznej inteligencji i robotyki w celu tłumienia zamieszek, protestów lub ruchów opozycji, z potencjalnie niepożądanymi konsekwencjami dla dynamiki politycznej.

W niniejszych rozważaniach głównym celem były długoterminowe konsekwencje otwartości. Jeśli weźmiemy pod uwagę długoterminowe konsekwencje, jednocześnie w funkcji oceny silnie uprzywilejowując wpływ na obecnie istniejących ludzi, to szczególnie istotną kwestią jest to, czy w ogóle otwarta tendencja ma wpływ na przyspieszenie rozwoju SI. Może to wynikać z dwóch powodów, po pierwsze szybszy rozwój SI oznaczałby szybsze wdrażanie blisko i średnioterminowych korzyści ekonomicznych wynikających z SI lub, co więcej, ponieważ szybszy rozwój SI zwiększałby prawdopodobieństwo, że niektórzy obecnie istniejący ludzie będą żyli wystarczająco długo, aby czerpać o wiele większe korzyści z superinteligencji maszyn, takie jak długowieczność i ekstremalny dobrobyt. Jeśli zamiast tego funkcja oceny nie uprzywilejowywałaby obecnych ludzi w stosunku do przyszłych pokoleń, szczególnie ważnym czynnikiem byłby wpływ otwartości na łączną sumę przyszłego ryzyka egzystencjalnego.

W kontekście, gdzie nacisk kładziony jest na skutki długoterminowe, a zwłaszcza na skumulowane ryzyko egzystencjalne, przedstawiono analizę dotyczącą dwóch kluczowych wyzwań: problemu kontroli i problemu politycznego. Zidentyfikowaliśmy trzy kategorie potencjalnego wpływu otwartości na te problemy. Argumentowaliśmy, że pierwsza z tych kategorii, wpływ otwartości na przyspieszenie rozwoju SI – jak się wydaje – ma stosunkowo słabe implikacje strategiczne. Nasza analiza koncentrowała się zatem głównie na dwóch pozostałych kategoriach: otwartości, dzięki której rozwój SI jest bardziej konkurencyjny, oraz otwartości umożliwiającej szersze zaangażowanie.

Uczynienie wyścigu rozwoju SI bardziej konkurencyjnym ma istotny negatywny wpływ na problem kontroli, zmniejszając zdolność wiodącej jednostki do wstrzymania prac lub zaakceptowania niższego poziomu wydajności w celu wprowadzenia środków kontroli. Może to przyczynić się do zwiększenia ryzyka egzystencjalnego związanego ze zmianami SI. Zintensyfikowana konkurencja może również wiązać się ze zwiększeniem prawdopodobieństwa konkurowania SI. Niemniej jednak strategiczny efekt netto może być niejasny i mieć mniejszą wagę decyzyjną niż efekt „brak opcji do spowolnienia". Istnieje również szereg implikacji związanych z intensywną konkurencją w rozwoju SI dla problemu politycznego: zmniejszenie prawdopodobieństwa zmonopolizowania zalet zaawansowanej SI przez małą grupę (atrakcyjne), zmniejszenie prawdopodobieństwa wystąpienia singletonu (co może być katastrofalne). Może też wystąpić pewne niejednoznaczne oddziaływanie na oczekiwany względny wpływ potencjału status quo na przyszłość post-SI, prawdopodobnie zwiększając ten wpływ w scenariuszach ograniczonych sprzętowo i zmniejszając go w scenariuszach ograniczonych programowo. Ponownie, z perspektywy minimalizacji ryzyka egzystencjalnego, oddziaływanie netto tych implikacji otwartości na problem polityczny wydaje się negatywne.

Otwartość umożliwiająca szersze zaangażowanie może mieć istotny pozytywny wpływ na problem kontroli, a mianowicie umożliwia zewnętrznym badaczom, którzy mogą być mniej stronniczy i bardziej zainteresowani bezpieczeństwem publicznym, pracować z najnowocześniejszymi systemami SI. Innym sposobem, w jaki otwartość może pozytywnie wpłynąć na problem kontroli, jest umożliwienie lepszego planowania społecznego i ustalania priorytetów. Korzyść ta nie wymagałaby otwartości na szczegółowe informacje techniczne, tylko na temat planów i możliwości dotyczących projektów SI. Jeżeli otwartość prowadzi do większego zaangażowania, może to mieć również wpływ na problem polityczny, umożliwiając lepsze przewidywanie i zwiększając tym samym prawdopodobieństwo rządowej kontroli nad zaawansowaną SI. To, czy wartość oczekiwana będzie dodatnia czy ujemna, nie jest całkowicie jasne. Może to zależeć na przykład od tego, kto kontrolowałby zaawansowaną SI w przypadku, gdyby nie została ona znacjonalizowana. Podsumowując, być może jednak można ocenić konsekwencje dla problemu politycznego szerokiej gamy podmiotów zyskujących zwiększoną zdolność przewidywania jako pozytywne. Ponownie można zauważyć, że odpowiednim rodzajem otwartości jest tutaj otwartość na możliwości, cele i plany, a nie otwartość na szczegóły techniczne i kod. Otwartość na szczegóły techniczne i kod mogą mieć mniejszy wpływ na ogólną dalekowzroczność i może zmniejszyć prawdopodobieństwo wywłaszczenia.

Specyficzne formy otwartości

Otwartość może przybierać różne formy: otwartość w zakresie nauki, kodu źródłowego, danych, technik bezpieczeństwa lub możliwości, oczekiwań, celów, planów i struktury za-rządzania projektem sztucznej inteligencji. W zakresie, w jakim możliwe jest otwarcie się w niektórych z tych wymiarów bez ujawnienia wielu informacji na temat innych wymiarów, można zadać bardziej szczegółowe pytanie polityczne, a odpowiedź może być różna dla różnych form otwartości.

Nauka ikod źródłowy

Otwartość na modele naukowe, algorytmy i kod źródłowy była przedmiotem większości poprzednich dyskusji. Jednym z dodatkowych niuansów jest to, że optymalna strategia może zależeć od czasu. Jeśli SI zaawansowanego rodzaju, dla której problem kontroli staje się krytyczny, jest dość odległa, może się zdarzyć, że wszelkie informacje, które zostaną ujawnione w wyniku bardziej otwartej polityki programistycznej, i tak będą szeroko rozpowszechnione przed osiągnięciem końcowego etapu. W takim przypadku wcześniejszy główny argument przeciwko otwartości nauki i kodu, mówiący o tym, że aby

rozwój SI był bardziej konkurencyjny i zmniejszyłby możliwość jej spowolnienia, może nie dotyczyć dzisiejszej otwartości. Może więc być możliwe czerpanie krótkoterminowych korzyści z otwartości, przy jednoczesnym uniknięciu kosztów długoterminowych, zakładając, że projekt może rozpocząć się jako otwarty, a następnie przejść do zamkniętej polityki rozwoju w odpowiednim czasie. Należy jednak pamiętać, że wprowadzenie w krytycznym czasie opcji zamknięcia usunie jedną z głównych przyczyn faworyzowania otwartości, a mianowicie nadzieję, że otwartość zmniejszy prawdopodobieństwo monopolizacji korzyści zaawansowanych SI. Jeśli polityka otwartości jest odwracalna, nie może służyć jako wiarygodne zobowiązanie do dzielenia się owocami zaawansowanej SI. Niemniej jednak nawet ludzie, którzy nie sprzyjają otwartości na późnych etapach, mogą sprzyjać otwartości na etapach wczesnych ze względu na niższe koszty otwartości.

Metody kontroli i analiza ryzyka

Otwartość na techniki bezpieczeństwa wydaje się jednoznacznie dobra, przynajmniej jeśli nie rozleje się zbytnio na inne formy otwartości. Należy zachęcać twórców SI do dzielenia się informacjami na temat potencjalnego ryzyka związanego z zaawansowaną SI oraz technikami kontrolowania takiej SI. Należy podjąć wysiłki, aby umożliwić zewnętrznym badaczom wniesienie wkładu pracy i niezależnych perspektyw w badania nad bezpieczeństwem, jeżeli można to zrobić bez ujawniania zbyt dużej ilości wrażliwych informacji.

Możliwości i oczekiwania

Otwartość na możliwości i oczekiwania co do przyszłych postępów, jak już przedstawiono, daje mieszany efekt, umożliwiając lepszy nadzór społeczny i dostosowanie, podczas gdy w niektórych modelach pojawia się ryzyko zaostrzenia dynamiki wyścigu. Niektórzy aktorzy mogą próbować kierować odkrycia do określonych odbiorców, którzy ich zdaniem będą szczególnie konstruktywni. Na przykład technokraci mogą martwić się tym, że szerokie zaangażowanie publiczne w kwestię zaawansowanej SI wygeneruje więcej zamieszania niż zysków, powołując się na analogiczne przypadki, takie jak debaty dotyczące GMO w Europie, gdzie mogłoby się wydawać, że korzystny postęp technologiczny mógłby być kontynuowany z mniejszą liczbą przeszkód, a rozmowa była bardziej zdominowana przez elity naukowe i polityczne przy małym zaangażowaniu opinii publicznej. Z kolei zwolennicy demokracji bezpośredniej mogą oponować, że omawiane kwestie były zbyt ważne, aby mogły zostać rozstrzygnięte przez grupę programistów SI, dyrektorów technicznych lub urzędników państwowych (którzy mogą służyć interesom parafialnym), oraz że społeczeństwo i świat lepiej funkcjonują przy szeroko otwartej dyskusji, która wyraża wiele różnorodnych poglądów i wartości.

Wartości, cele i struktury zarządzania

Otwartość na wartości, cele i struktury zarządzania jest ogólnie mile widziana, ponieważ powinna dążyć do różnicowego wspierania projektów ukierunkowanych na cele atrakcyjne dla szerokiego grona zainteresowanych stron. Otwartość w tych kwestiach może również sprzyjać zaufaniu oraz zmniejszać presję na to, by poświęcać bezpieczeństwo ze względu na przewagę konkurencyjną. Im bardziej konkurenci czują, że nadal będą mogli czerpać zyski z sukcesu rywala, tym większe są szanse na wspólne podejście lub przynajmniej takie, w którym konkurenci nie działają aktywnie przeciwko sobie. Z tego powodu pożądane są środki, które wyrównują motywacje pomiędzy różnymi twórcami SI, szczególnie motywacje na późniejszych etapach. Takie środki mogą obejmować wzajemne utrzymywanie zasobów, wspólne przedsięwzięcia badawcze, formalne lub nieformalne zobowiązania do współpracy, poparcie zasad stwierdzających, że zaawansowana SI powinna być rozwijana tylko dla wspólnego dobra oraz inne działania, które budują zaufanie i przyjaźń między zwolennikami.

PODZIĘKOWANIE

Prace te zostały wsparte przez Europejską Radę ds. Badań Naukowych. Jestem wdzięczny za pomocne komentarze i dyskusje następującym osobom: Stuartowi Armstrongowi, Owenowi Cotton-Barrattowi, Robowi Bensingerowi, Paulowi Christiano, Allanowi Dafoe, Ericowi Drexlerowi, Owainowi Evansowi, Oliverowi Habryka, Demisowi Hassabisowi, Shane'owi Leggowi, Javierowi Lezaunowi, Luke Muehlhauserowi, Toby'emu Ordowi, Guy-owi Ravine'owi, Steve'owi Raynerowi,

⇨ Redakcja: dr Roman V. Yampolskiy
Bibliografia dostępna pod linkiem: nis.com.pl/bibliografia.html