DOI: 10.24425/academiaPAN.2024.152334

PEXELS – URFRIENDLYPHOTOG

Od sztucznej inteligencji do gwiazd

Rozwój naukowo-techniczny nie jest neutralny ideologicznie. Napędza go nie tylko pragnienie poznania rzeczywistości, lecz także chęć zrealizowania za pomocą nowych technologii konkretnych wizji ideowych.

Filip Biały

Uniwersytet im. Adama Mickiewicza w Poznaniu
Uniwersytet w Manchesterze

Dziś wybitnie zideologizowane zdają się badania nad sztuczną inteligencją, w której reprezentanci tescrealizmu – nowej filozofii Doliny Krzemowej – upatrują szansy na dokonanie przez ludzkość skoku ewolucyjnego.

„Z San Francisco najszybciej można dostrzec przyszłość" – tak brzmi pierwsze zdanie eseju *Situational Awareness* z 2024 roku. Jego autor, Leopold Aschenbrenner, były pracownik OpenAI, przekonuje, że już w 2027 roku dojdzie do przełomu w badaniach nad sztuczną inteligencją (SI). Opierając się na ekstrapolacji trendów z ostatnich kilku lat, powinniśmy spodziewać się, że SI osiągnie wkrótce poziom ponadludzki. „Nie trzeba wierzyć w science fiction, wystarczy wierzyć w prostą linię na wykresie" – pisze Aschenbrenner, skrupulatnie dołączając rzeczony wykres, na którym jest przedstawiony rosnący liniowo poziom zdolności poznawczych kolejnych wersji modelu językowego GPT. Jeżeli w ciągu czterech lat udało się dojść od modelu GPT-2 (poziom przedszkolaka) do GPT-4 (poziom mądrego licealisty), to w 2027 roku model ten osiągnie poziom osoby z doktoratem albo „automatycznego badacza/inżyniera SI".

Debata wokół SI jest pełna podobnych enuncjacji, zapewniających, że wiedzę o przełomowej technice posiada co najwyżej kilkaset osób, pracujących

dr Filip Biały
Pracownik naukowy Uniwersytetu im. Adama Mickiewicza w Poznaniu oraz Uniwersytetu w Manchesterze, wykładowca European New School of Digital Studies. Jego badania skupiają się na konsekwencjach politycznych procesów transformacji cyfrowej.
filip.bialy@amu.edu.pl

w OpenAI czy innym z laboratoriów SI. Autorów tych rewelacji – zdecydowana większość z nich to mężczyźni – charakteryzuje quasi-religijna wiara w rychłe nadejście tzw. ogólnej sztucznej inteligencji (ang. *artificial general intelligence* – AGI) bądź superinteligencji (ang. *artificial superintelligence* – ASI). Mowa o technice zdolnej wykonywać wszelkie zadania na poziomie ludzkim bądź wyższym. Jak zobaczymy, intensywne prace nad AGI/ASI są jednak zarówno przedmiotem nadziei, jak i zasadniczych obaw.

Ścieżka do transhumanizmu

Jeszcze dekadę temu liderom technologicznym z okolic San Francisco przypisywano tzw. ideologię kalifornijską, łączącą kontrkulturowe wartości hippisów z kalkulacją finansową kapitalistycznych yuppies. Dziś Elon Musk, Mark Zuckerberg, Sam Altman i wielu innych pozostają pod wpływem innego koktajlu ideologicznego. Określa go zaproponowany w 2023 roku przez Timnit Gebru i Émile'a P. Torresa akronim TESCREAL. Składają się na niego nietożsame, lecz bliskie sobie i nierzadko wyznawane łącznie przez te same osoby nurty ideowe: transhumanizm, ekstropianizm, syngularytyzm, kosmizm, racjonalizm, efektywny altruizm i longtermizm.

> Przez osobliwość technologiczną można także rozumieć osiągnięcie przez SI poziomu, który doprowadzi do „eksplozji inteligencji".

Choć wymienione określenia mogą się wydawać dziwacznymi neologizmami, odnoszą się do idei rozwijanych od wielu dekad. Ich wspólnym genealogicznym poprzednikiem jest eugenika. W początkach XX wieku eugenicy mówili o metodach selektywnego rozmnażania się ludzi, by ulepszać kolejne pokolenia. Wraz z nadejściem inżynierii genetycznej eugenika zmieniła swój charakter, odcinając się od rasistowskich korzeni, a skupiając się na możliwości eliminacji wad i chorób genetycznych. Eugenika trzeciej generacji, istota tescrealizmu, jest jeszcze ambitniejsza. Chodzi o przejście na kolejne stadium ewolucyjne, którym ma być digitalizacja ludzkiej świadomości, a następnie ekspansja cyfrowych ludzi we wszechświecie.

Tescrealizm nie jest jednolitym ruchem i prawie nikt nie nazwałby siebie tescrealistą. A jednak przyglądając się temu, kto wyznaje idee należące do tej rodziny ideologicznej, dostrzeżemy powtarzające się nazwiska, a także wiele powiązań instytucjonalnych i politycznych.

Akronim TESCREAL w zamierzeniu Timnit Gebru i Emile'a P. Torresa, twórców tego pojęcia, odzwierciedla chronologiczne pojawianie się kolejnych nurtów. Pierwszy był transhumanizm. Pojęcie to zdefiniował w 1957 roku Julian Huxley, jednak jego obecne rozumienie zawdzięczamy filozofowi Maxowi More'owi. Pisał on w końcu lat 80. XX wieku o dążeniu do kondycji postludzkiej: radykalnej zmiany naszej natury przy wykorzystaniu osiągnięć nauki i techniki.

More był współzałożycielem Extropy Institute, emanacji ekstropianizmu. Nurt ten głosi dążenie do osiągnięcia nieśmiertelności i nieograniczonej ekspansji przez użycie inteligentnej techniki. Wiara w zaistnienie niezbędnego przełomu technicznego łączy ekstropian z syngularytystami, takimi jak Ray Kurzweil. Kurzweil upowszechnił 20 lat temu koncepcję osobliwości technologicznej (*singularity*), przez którą rozumie połączenie człowieka ze SI. W książce *The Singularity is Nearer* (2024) Kurzweil powtarza swoje przewidywania, wskazując 2029 rok jako moment osiągnięcia przez SI poziomu ludzkiej inteligencji oraz 2045 rok jako perspektywę integracji ludzi z maszynami.

Połączenie SI z człowiekiem jest jedną z centralnych idei kosmizmu, którego antecedencje sięgają filozofii rosyjskiej z przełomu XIX i XX wieku. Dziś kosmizm jest reprezentowany m.in. przez Bena Goertzela, który spopularyzował pojęcie ogólnej sztucznej inteligencji (AGI). Elementami jego wizji jest możliwość załadowania ludzkiego umysłu do komputera oraz kolonizacja przestrzeni kosmicznej.

Przez osobliwość technologiczną można także rozumieć osiągnięcie przez SI poziomu, który doprowadzi do „eksplozji inteligencji". Zamiast podporządkować się człowiekowi, zacznie ona tworzyć doskonalsze wersje samej siebie. Koncepcję tę współcześnie opisał filozof Nick Bostrom, autor wpływowej *Superinteligencji* z 2014 roku. Zagłady grożącej ze strony SI obawia się Eliezer Yudkowsky, określany z tego względu mianem doomera. Yudkowsky stworzył platformę LessWrong, skupiającą przedstawicieli tescrealistycznej wersji racjonalizmu, dążącego do ulepszenia ludzkiego rozumowania i decydowania. Ma temu służyć zaawansowana SI, pod warunkiem że nie wymknie się spod kontroli, co jest największą troską Yudkowsky'ego i założonego przez niego Machine Intelligence Research Insitute.

Racjonalistom blisko do dwóch ostatnich nurtów tescrealizmu: efektywnego altruizmu i longtermizmu. Efektywny altruizm (EA) jest najnowszą, doskonale zorganizowaną odsłoną etyki utylitarystycznej. Filozoficznymi ojcami EA byli Peter Singer oraz Derek Parfit, uczący, że priorytetowe traktowanie w naszych decyzjach etycznych ludzi przestrzennie bądź czasowo nam bliskich jest błędne. Powinniśmy

stosować uniwersalne reguły wobec każdego człowieka, nawet tego, który znajduje się na drugim końcu świata, bądź – tu EA łączy się z longtermizmem – jeśli człowiek ten narodzi się miliony lat po nas. Współtwórca EA, William MacAskill, pisał w książce *What Do We Owe the Future* (2022), że ponieważ w perspektywie milionów lat ludzkość, która zasiedli przestrzeń kosmiczną, będzie wielokrotnie liczniejsza od współczesnej (wedle wyliczeń Bostroma ma to być nawet 10^{58} istnień), te właśnie przyszłe pokolenia muszą być podstawowym czynnikiem w naszych kalkulacjach etycznych.

W odniesieniu do żyjących tu i teraz EA nawołuje do wyboru takich ścieżek kariery, które będą wieść do osiągnięcia szczęścia jak największej liczby przyszłych ludzi. Prowadzi to pozornie do paradoksów, takich jak wyższe cenienie tych, którzy zajmują się rozwojem SI, od tych, którzy chcieliby zapobiec katastrofie klimatycznej. Jest tak dlatego, że EA i longtermizm opierają się na kategorii ryzyka egzystencjalnego. Chodzi o takie wydarzenia, które mogłyby uniemożliwić zrealizowanie długofalowych wizji podboju kosmosu przez zdigitalizowaną ludzkość. Kryzys klimatyczny może być dla miliardów dramatyczny w skutkach, jednak tak jak pandemie, wojny czy brak sprawiedliwości społecznej najprawdopodobniej nie doprowadzi do całkowitej zagłady człowieka. Dlatego też nie są to dla EA problemy priorytetowe.

Elegia dla bogaczy

Pierwsze przykazanie EA brzmi: „Zarabiaj, by dawać". Jego wyznawcą był sponsor idei EA i bliski znajomy MacAskilla, Sam Bankman-Fried, skazany za defraudację miliardów dolarów z założonej przez siebie giełdy kryptowalut FTX. Choć może wydawać się niesprawiedliwe, by idee EA rozliczać skrajnym przykładem oszusta finansowego, przypadek ten doskonale obrazuje przenikanie się ideologii tescrealistycznych oraz praktyk korporacyjnych Doliny Krzemowej.

Gdyby tescrealizm był bowiem jedynie osobliwą filozofią garstki ekscentrycznych entuzjastów postępu technicznego, można by pewnie potraktować go jako intelektualny odprysk epoki cyfryzacji. Jest to dziś jednak środowisko wpływowe, bardzo dobrze zinstytucjonalizowane, a jeszcze lepiej finansowane. Sam tylko Future Fund założony przez Bankmana-Frieda przeznaczył 160 mln dolarów na cele wspierane przez EA.

Tescrealiści uważają, że najlepiej rokującym środkiem realizacji ich wizji jest ogólna SI. Za główne ryzyko egzystencjalne uznają pojawienie się autonomicznej SI niedopasowanej do ludzkich wartości. Jeśli będzie ona dążyć do maksymalizacji ustalonych przez siebie celów, czynnik ludzki może zostać uznany za przeszkodę, którą trzeba wyeliminować. Dlatego tak ważne są prace nad bezpieczeństwem sztucznej inteligencji, czyli nad metodami dopasowania SI do pożądanych wartości. W tej perspektywie staje się zrozumiałe, dlaczego ci sami ludzie, tacy jak Elon Musk, jedną ręką podpisują wezwania do moratorium na badania nad SI, drugą zaś hojnie wspierają badania nad rozwojem „bezpiecznej" ogólnej SI.

Na pierwszy rzut oka ten aspekt tescrealizmu wydaje się pozytywny. Chyba dobrze, że tak dużą wagę przywiązuje on do bezpiecznego rozwoju techniki? Problem w tym, że przeznaczając miliardy na bezpieczeństwo wciąż nieistniejącej i być może niemożliwej do zbudowania ogólnej SI, nie kieruje się podobnych środków na zapobieganie szkodom wyrządzanym przez już działające systemy wąskiej SI. A to właśnie zastosowanie SI w decydowaniu o przyznawaniu świadczeń socjalnych, w prowadzeniu działań zbrojnych, w kontroli i inwigilacji mniejszości etnicznych i rasowych są palącym problemami ludzi żyjących dziś na Ziemi, a nie przyszłych, międzygwiezdnych pokoleń.

Rolą ideologii jest dostarczanie swoim wyznawcom pojęć strukturyzujących świat społeczny, motywującej wizji przyszłości oraz praktycznych uzasadnień podejmowanych działań. Nie ulega wątpliwości, że tescrealizm taką rolę odgrywa, wkraczając coraz wyraźniej w główny nurt polityki. Jeśli ktoś w to wątpi, powinien przyjrzeć się uważniej drodze politycznej J.D. Vance'a, którego Donald Trump wybrał jako kandydata na wiceprezydenta.

J.D. Vance został senatorem dzięki finansowemu wsparciu udzielonemu przez Petera Thiela, sponsora EA i innych inicjatyw tescrealistycznych. Thiel, jeden z najważniejszych technologicznych *venture capitalists*, poznał Vance'a w 2011 roku, kilka lat później dał mu pracę w jednym ze swoich funduszy inwestycyjnych. Donald Trump przekonał się do Vance'a po serii rozmów z Muskiem i Thielem.

Po co tescrealistom Vance? Jak pisał „The Washington Post", chodzi o to, by w Białym Domu znalazł się ktoś, kto rozumie, że rozwijanie nowych technologii nie jest rolą rządu, jak było to w czasach Projektu Manhattan, lecz należy w tym względzie pozostawić wolną rękę „geniuszom" z Doliny Krzemowej.

Być może zatem ostatecznie tescrealizm jest przede wszystkim narzędziem realizacji ekonomicznego interesu miliarderów z Doliny Krzemowej, dostarczając obietnicy rozwiązania wszelkich ludzkich problemów nieskrępowanym rozwojem technicznym. Kiedy przestanie być do tego przydatny, podzieli losy klasycznej ideologii kalifornijskiej i zostanie zastąpiony inną filozofią. Zanim to jednak nastąpi, tescrealizm musi stać się przedmiotem krytycznej uwagi nie tylko badaczy ideologii, lecz także opinii publicznej. Szczególnie jeśli pragniemy, by rozwój techniczny był poddany demokratycznej kontroli i dopasowany do wartości sprzyjających dobrobytowi całych społeczeństw, a nie tylko korzyści wąskich elit kapitalizmu cyfrowego. ■