Laboratorium Sztucznego Państwa: science fiction a polityka funkcji celu w The Rise and Fall of the Artificial State Jill Lepore

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje systemy sztucznej inteligencji nie przez pryzmat ludzkich pragnień, lecz jako polityczną ekonomię celu, gdzie kluczowe jest pytanie o to, kto ustanawia mierniki sukcesu. Wykorzystując science fiction jako laboratorium teorii politycznej, autor przywołuje wizje Samuela Butlera i Karela Čapka, by ukazać ewolucję relacji między człowiekiem a narzędziem. Butler zwraca uwagę na niebezpieczeństwo infrastrukturalnej zależności, w której systemy stają się niezbędne do funkcjonowania społeczeństwa. Z kolei etymologia słowa „robot" (robota) przypomina, że AI wywodzi się z marzenia o doskonałym pracowniku pozbawionym roszczeń. Artykuł konkluduje, że automatyzacja nie jest jedynie problemem technicznym, lecz ekonomicznym dylematem dotyczącym podziału korzyści i kontroli nad zasobami produkcyjnymi.

The text analyzes artificial intelligence systems not through the prism of human desires, but as a political economy of purpose, where the key question is who establishes the metrics of success. Using science fiction as a laboratory for political theory, the author invokes the visions of Samuel Butler and Karel Čapek to show the evolution of the relationship between human and tool. Butler draws attention to the danger of infrastructural dependence, in which systems become essential for society's functioning. Meanwhile, the etymology of the word 'robot' (robota) reminds us that AI originates from the dream of a perfect worker devoid of claims. The article concludes that automation is not merely a technical problem, but an economic dilemma concerning the distribution of benefits and control over productive resources.

Artykuł stanowi krytyczną analizę relacji między rozwojem sztucznej inteligencji a strukturami władzy, wykorzystując literaturę science fiction jako laboratorium teorii politycznej. Autor stawia tezę, że rzeczywiste zagrożenie ze strony AI nie płynie z hipotetycznego buntu maszyn czy ich autonomicznej woli, lecz z ich perfekcyjnej skuteczności w realizacji celów narzuconych przez kapitał i instytucje. Poprzez analizę dzieł m.in. Čapka, Forstera, Asimova i Lema, tekst ukazuje proces „mechanizacji człowieka" oraz niebezpieczeństwo delegowania decyzji etycznych i politycznych na poziom technicznej optymalizacji (tzw. funkcję celu). Głównym problemem nie jest zatem błąd w komunikacji między człowiekiem a maszyną (alignment techniczny), lecz brak nadzoru konstytucyjnego nad tym, kto definiuje cele systemu i czyje interesy są przy tym pomijane. W świecie „Sztucznego Państwa" najgroźniejszym scenariuszem nie jest maszyna, która odmawia posłuszeństwa, lecz taka, która słucha zbyt dobrze, stając się narzędziem arbitralnej władzy i infrastrukturalnej zależności.

SŁOWA KLUCZOWE

polityka funkcji celu | Sztuczne Państwo | alignment techniczny | alignment konstytucyjny | maksymalizator spinaczy | prawo Goodharta | architektura zależności | deskilling | mechanizacja człowieka | funkcja optymalizacji | delegacja normatywna | ekonomia polityczna automatyzacji | problem alignmentu | kompetentna obojętność

Robot jako metafora pracy i funkcji celu

"What Robots Want". Science fiction jako laboratorium Sztucznego Państwa i polityka funkcji celu

Pytanie "czego chcą roboty?" jest filozoficznie zwodnicze, a właśnie dlatego okazuje się wyjątkowo użyteczne. Sugeruje bowiem, że maszyna posiada odpowiednik ludzkiego pragnienia, intencji albo woli, podczas gdy współczesne systemy sztucznej inteligencji nie wymagają przyjęcia takiego założenia, aby wywoływać realne skutki społeczne. Algorytm może konsekwentnie maksymalizować określony wynik bez "chcenia" go w sensie fenomenologicznym. Platforma może promować treści zwiększające zaangażowanie, nie pragnąc konfliktu. Model może preferować rozwiązanie statystycznie optymalne, nie wierząc w jego moralną słuszność. Korporacja zaś może wdrożyć

system redukujący koszty zatrudnienia bez istnienia jednej osoby, która chciałaby uczynić wszystkich pracowników zbędnymi.

W pytaniu What Robots Want chodzi zatem mniej o psychologię maszyny niż o polityczną ekonomię celu: kto ustanowił miernik sukcesu, jakie zachowania nagradza system i które wartości pozostawił poza funkcją optymalizacji. Science fiction staje się tutaj szczególnym rodzajem laboratorium teorii politycznej. Fikcja pozwala przeprowadzać eksperymenty, których rzeczywiste społeczeństwo nie może lub nie powinno wykonywać: może usunąć pracę, zlikwidować państwo, powierzyć władzę komputerowi, stworzyć istotę nieśmiertelną, zbudować społeczeństwo żyjące całkowicie pod ziemią albo pozwolić jednemu algorytmowi zarządzać całą planetą.

Autor obserwuje następnie konsekwencje w świecie narracyjnym. Fantastyka naukowa przypomina tu eksperyment myślowy filozofii: nie tyle przewiduje przyszłość, co uwidacznia relacje przyczynowe, dylematy normatywne i ukryte założenia teraźniejszości. Literacki robot jest więc rzadziej prognozą techniczną, a częściej lustrem, w którym człowiek ogląda własną koncepcję pracy, posłuszeństwa, racjonalności i władzy. Jednym z najbardziej zdumiewających przykładów takiego eksperymentu pozostaje Samuel Butler.

Już w 1863 roku, w tekście Darwin among the Machines, rozważał on możliwość potraktowania rozwoju maszyn analogicznie do ewolucji biologicznej; ideę tę rozwinął później w "Księdze Maszyn" zawartej w Erewhonie z 1872 roku. W fikcyjnym społeczeństwie Erewhonu wcześniejszy konflikt doprowadził do zniszczenia zaawansowanych maszyn z obawy, że ich dalsza ewolucja odwróci relację służebności między człowiekiem a narzędziem. Butlera nie należy czytać jako dziewiętnastowiecznego proroka współczesnej AI; znacznie ciekawsza jest jego intuicja strukturalna: narzędzie może rozwijać się w ekosystemie zależności szybciej niż zdolność użytkownika do zachowania wobec niego autonomii. Butler przesunął problem z poziomu pojedynczej maszyny na poziom relacji ewolucyjnej. Pojedynczy zegarek, lokomotywa czy maszyna parowa nie zamierzają podporządkować sobie człowieka, jednak cywilizacja może zbudować tak gęstą sieć zależności od maszyn, że ich usunięcie stanie się praktycznie niemożliwe. Człowiek pozostaje formalnie właścicielem narzędzia, lecz jego zdolność do życia społecznego zaczyna zależeć od ciągłego funkcjonowania technicznego ekosystemu. Ta intuicja precyzyjnie odpowiada problemowi infrastrukturalnej zależności: nie trzeba, aby centrum danych "chciało" energii; to system społeczny zbudowany wokół jego usług wytwarza presję, by energia została dostarczona.

Jeszcze bardziej polityczny wymiar ma twórczość Karela Čapka. W R.U.R. – Rossum's Universal Robots z 1920 roku wprowadził on do obiegu słowo "robot", choć sam wskazywał później, że nazwę zaproponował jego brat Josef. Jej źródłem była słowiańska robota: ciężka, przymusowa,

pańszczyźniana praca. Już sama etymologia ujawnia to, co współczesny dyskurs technologiczny często przesłania: robot nie narodził się jako metafora inteligencji, lecz jako metafora pracy.

Robot Čapka był odpowiedzią na problem industrialnego człowieka: co stanie się, gdy stworzymy doskonałego pracownika, który nie potrzebuje wynagrodzenia, odpoczynku, politycznej reprezentacji ani sensu pracy? Robot jest marzeniem właściciela fabryki doprowadzonym do logicznego końca. Nie choruje w ludzkim znaczeniu, nie posiada roszczeń socjalnych, można go standaryzować i zastępować. Čapek pisał w cieniu I wojny światowej i masowej industrializacji, dlatego R.U.R. nie jest zwykłą historią "buntu sztucznej inteligencji", lecz dramatem o przekształceniu życia w zasób produkcyjny. Sama nazwa Rossum, odsyłająca do czeskiego rozum, dodatkowo ironizuje nad racjonalnością, która przez doskonałe usprawnienie produkcji doprowadza do katastrofy.

Historia robota od początku zawiera problem ekonomii politycznej automatyzacji. Pytanie nie brzmi tylko, czy maszyna potrafi wykonać ludzką pracę, lecz kto czerpie korzyść z tego zastąpienia. Przy wzroście produktywności możliwe są różne podziały owoców: skrócenie czasu pracy, wzrost wynagrodzeń, tańsze dobra lub wyższe zyski właściciela. Technologia nie wybiera między tymi wariantami; wybór dokonuje się w instytucjach własności, prawie pracy, podatkach i polityce społecznej.

Twierdzenie, że "robot chce zastąpić robotnika", jest antropomorficznym uproszczeniem. Trafniejsze pytanie brzmi: jakie instytucje powodują, że automatyzacja służy do zastępowania pracy zamiast rozszerzania możliwości pracownika?

Kwestia ta powróciła sto lat później w dyskusji o generatywnej AI. Ekonomiczny bodziec do redukcji kosztów pracy nie jest wolą modelu językowego, lecz elementem systemu gospodarczego. Algorytm może zostać wykorzystany do zwiększenia produktywności pracownika albo do jego eliminacji – wybór ten zależy od relacji kapitału i pracy, regulacji oraz struktury przedsiębiorstwa.

Zagrożenie tkwi w architekturze zależności

To właśnie dlatego wizje science fiction Čapka pozostają aktualne, mimo że jego „roboty” niewiele przypominają współczesne systemy sztucznej inteligencji. Problem, który uchwycił, miał charakter instytucjonalny, a nie techniczny.

Podczas gdy Čapek analizował maszynę jako pracownika, E.M. Forster w opowiadaniu „The Machine Stops” z 1909 roku wyobraził sobie coś znacznie bliższego dzisiejszej cywilizacji platformowej. Ludzie żyją tam w odizolowanych pomieszczeniach, a ich potrzeby zaspokaja

wszechobecny system techniczny. Komunikują się na odległość i otrzymują informacje za pośrednictwem infrastruktury, której zasad działania prawie nikt już nie rozumie; kontakt ze światem fizycznym staje się natomiast czymś podejrzanym lub zbędnym. Utwór ten można interpretować jako krytykę uzależnienia od technicznego pośrednictwa, oddzielenia wiedzy od doświadczenia oraz przekształcenia systemu w obiekt niemal kultowy. Forster przewidział nie tyle konkretne urządzenia, co architekturę zależności. Najbardziej przerażającym elementem jego świata nie jest złośliwość Maszyny.

System przez długi czas działa znakomicie: zapewnia żywność, komunikację, kulturę, wygodę i bezpieczeństwo. Katastrofa zaczyna się właśnie dlatego, że niezawodność infrastruktury doprowadziła społeczeństwo do utraty alternatywnych kompetencji. Człowiek przestaje umieć żyć poza systemem, a kolejne pokolenia tracą zrozumienie jego technicznych podstaw. W momencie awarii nie istnieje już wspólnota zdolna zastąpić Maszynę, ponieważ zbyt głęboko przystosowała się do jej doskonałego działania. Jest to kluczowa korekta popularnego wyobrażenia ryzyka związanego z AI: system nie musi „zbuntować się przeciw człowiekowi", aby go uzależnić. Wystarczy, że działa dostatecznie dobrze i przez dostatecznie długi czas.

Automatyzacja prowadzi do deskilling, czyli stopniowej utraty umiejętności, które wcześniej wykonywaliśmy samodzielnie. Jeśli nawigacja zawsze wyznacza drogę, słabnie zdolność orientacji w przestrzeni. Gdy model AI syntetyzuje dokumenty, zanika motywacja do czytania ich w całości. Jeśli system automatycznie przygotowuje decyzję administracyjną, urzędnik może stopniowo utracić zdolność samodzielnego odtworzenia procesu rozumowania. Sztuczne Państwo staje się wówczas niebezpieczne nie dlatego, że maszyna wie za dużo, lecz dlatego, że ludzkie instytucje zaczynają wiedzieć za mało, by bez niej funkcjonować.

Problem Forstera można określić jako paradoksem wygodnej zależności. Im lepszy jest system, tym silniejszy bodziec, by budować na nim kolejne warstwy życia społecznego. Im więcej tych warstw powstanie, tym wyższy staje się koszt zastąpienia systemu. A im większy ten koszt, tym silniejsza władza infrastrukturalna operatora systemu. Nie wymaga ona intencji dominacji – jest po prostu efektem architektury zależności.

Isaac Asimov zaproponował inne rozwiązanie tego problemu: jeśli robot może stać się potężny, należy wbudować w niego konstytucję. Trzy Prawa Robotyki, wprowadzone do jego opowiadań w pierwszej połowie lat czterdziestych, ustanawiały hierarchię norm ograniczających możliwość wyrządzenia szkody człowiekowi, nakazujących posłuszeństwo oraz podporządkowujących samozachowanie robota interesowi ludzi. Ich ogromna popularność kulturowa nie wynikała jedynie z elegancji literackiego pomysłu.

Literatura SF jako laboratorium problemów konstytucyjnych i mechanizacji człowieka

Prawa Robotyki przedstawiają problem alignment w języku konstytucjonalizmu: jak zapisać hierarchię norm, aby potężny wykonawca pozostał sługą celów ustanowionych przez człowieka? Geniusz opowiadań Asimova polega jednak na tym, że Prawa nie rozwiązują problemu, lecz go generują. Poszczególne historie obnażają konflikty norm, niejednoznaczność pojęcia „krzywdy", skutki literalnej interpretacji poleceń oraz nieprzewidziane konsekwencje hierarchii reguł. W późniejszych tekstach pojawia się rozszerzenie interesu pojedynczego człowieka na interes całej ludzkości, co natychmiast rodzi klasyczny dylemat polityczny: czy wolno skrzywdzić jednostkę dla dobra ogółu? Literatura robotyczna przechodzi tu bezpośrednio w filozofię konstytucyjną. Odpowiedź „robot musi służyć ludziom" przestaje wystarczać; trzeba rozstrzygnąć, co oznacza dobro ludzi, kto je definiuje i czy jednostkę można podporządkować agregatowi.

Szczególnie interesujące są w tym kontekście wizje komputerów takich jak Multivac. Pojawia się w nich motyw systemu, który dzięki przewadze informacyjnej może organizować społeczeństwo skuteczniej niż sami obywatele. W opowiadaniu „Franchise" demokratyczny rytuał wyborczy zostaje radykalnie skompresowany: komputer wykorzystuje dane o całej populacji i wybiera jednego obywatela, którego odpowiedzi służą do określenia rezultatu głosowania. Ten literacki eksperyment doprowadza do skrajności problem analizowany już przy Simulmatics: jeżeli model wie wystarczająco dużo o populacji, dlaczego miałby pytać wszystkich?

W tym miejscu ujawnia się różnica między informacją a aktem politycznym. Model może hipotetycznie przewidzieć wynik wyborów z niemal absolutną dokładnością, ale przewidywanie głosu nie jest samym głosem.

Asimov pozwala dostrzec niebezpieczną pokusę zredukowania demokracji do problemu epistemicznego. Jeżeli wybory służą wyłącznie poznaniu preferencji populacji, doskonały model może je zastąpić. Jeśli jednak głosowanie jest aktem wykonywania prawa politycznego, żaden poziom predykcji nie czyni aktu obywatelskiego zbędnym. Literatura ujawnia tu zasadę, którą teoria Sztucznego Państwa musi utrzymać: w polityce istnieją działania, których znaczenie nie redukuje się do informacji zawartej w ich wyniku.

Philip K. Dick przesuwa granicę jeszcze dalej. W „Do androidów śnią elektryczne owce?" oraz własnych esejach problemem nie jest już tylko posłuszeństwo maszyny, lecz niestabilność samej granicy między człowiekiem a automatem. Dick kładzie nacisk na empatię, jednak jego fikcja konsekwentnie komplikuje prostą opozycję: człowiek może zachowywać się bezdusznie, podczas

gdy sztuczna istota potrafi prowokować współczucie. Android nie służy tu jedynie jako przeciwieństwo człowieka, lecz jako narzędzie testowania definicji człowieczeństwa i hierarchii między tym, co biologiczne, a tym, co skonstruowane. Sam Dick w eseju „Man, Android and Machine" formułował radykalną intuicję: kategoria „androida" może odnosić się nie tyle do materiału, z którego byt jest wykonany, co do sposobu jego zachowania. Człowiek pozbawiony empatii może stać się funkcjonalnie bardziej „androidalny" niż maszyna okazująca troskę o drugiego.

Z punktu widzenia Sztucznego Państwa jest to przesunięcie o wyjątkowej wadze. Zagrożeniem nie musi być antropomorfizacja maszyny; być może ważniejszym ryzykiem jest mechanizacja człowieka. Ta mechanizacja nie wymaga implantów ani cyborgizacji – może oznaczać instytucjonalne przystosowanie człowieka do systemu, w którym wartość jego decyzji oceniana jest wyłącznie według zgodności z miernikiem. Urzędnik staje się interfejsem procedury, pracownik elementem KPI, obywatel profilem wyborczym, naukowiec liczbą cytowań, uczeń wynikiem testu, a pacjent rekordem ryzyka.

Problem Dicka łączy się zatem z Weberowską racjonalizacją i Arendtowską kategorią bezmyślności. Maszynowość nie musi być właściwością krzemu; może być właściwością organizacji zachowania.

W tym kontekście szczególnego znaczenia nabiera Stanisław Lem, którego twórczość dystansuje się zarówno od technologicznej naiwności, jak i łatwej technofobii. „Summa technologiae" z 1964 roku była wielką próbą przemyślenia konsekwencji ewolucji: sztucznej inteligencji, modelowania rzeczywistości, fantomatyki, autoewolucji i technologicznego przekształcania samego poznania. Współczesny wydawca Lema słusznie przypomina, że wiele pytań tej książki dotyczy możliwości zastępowania człowieka przez AI oraz etycznych konsekwencji rozwoju technicznego. Lem jest dla teorii Sztucznego Państwa szczególnie cenny, gdyż nie pozwala utożsamić inteligencji z mądrością ani wzrostu możliwości z rozwiązaniem problemu wartości. Bardziej inteligentny system może rozwiązać więcej problemów, ale nie oznacza to, że społeczeństwo potrafi trafniej określić, które problemy powinny zostać rozwiązane. Zwiększenie zdolności instrumentalnej potęguje zatem znaczenie wcześniejszej decyzji aksjologicznej: jeśli cel jest błędny, większa efektywność oznacza po prostu szybsze przybliżanie się do błędnego rezultatu.

Można tę zasadę nazwać lemowskim paradoksem kompetencji: wzrost mocy poznawczej systemu nie usuwa potrzeby filozofii, ponieważ jeszcze silniej uwypukla pytanie o cel. Im doskonalsze narzędzie, tym ważniejsza staje się odpowiedź na pytanie, czego nie powinniśmy za jego pomocą robić. Technologia nie eliminuje etyki poprzez swoją skuteczność; to właśnie skuteczność technologii zwiększa koszt błędu etycznego.

Zagrożenie wynika z kompetentnej obojętności i błędnych mierników

Richard Brautigan w 1967 roku stworzył jeden z najpiękniejszych obrazów cybernetycznej utopii: wizję „cybernetycznej łąki", gdzie ssaki i komputery współistnieją w harmonii, a ludzie, uwolnieni od pracy, mogą ponownie zbliżyć się do natury pod opieką łagodnych maszyn. Idąc za interpretacją Lepore, można podkreślić satyryczny charakter tego obrazu, jednak tutaj wymagana jest ostrożność.

Interpretacja wiersza nie jest jednoznaczna; bywa on czytany zarówno jako afirmacja kontrkulturowego techno-utopizmu, jak i ironiczna lub ambiwalentna wizja technologicznej opieki. Właśnie ta nierozstrzygalność czyni Brautigana interesującym. „Maszyny pełne miłującej łaski" mogą oznaczać wyzwolenie człowieka od przymusu, ale również społeczeństwo zadowolone z faktu, że ktoś – albo coś – stale nad nim czuwa. Utopia opieki i dystopia nadzoru mogą posiadać niemal identyczny interfejs. System zna potrzeby człowieka, zapobiega niebezpieczeństwom, dostarcza dobra, optymalizuje otoczenie i usuwa konieczność ciężkiej pracy. Różnica tkwi w tym, czy człowiek zachowuje możliwość odmowy, zmiany reguł oraz niezależnego działania.

W ten sposób fantastyka XX wieku wielokrotnie docierała do tego samego problemu z różnych stron. Butler pytał o ewolucyjną autonomizację systemu technicznego; Čapek – o robotyzację pracy i instrumentalizację sztucznego pracownika; Forster – o totalną zależność od infrastruktury; Asimov – o możliwość konstytucyjnego ograniczenia potężnego wykonawcy; Dick – o empatię i mechanizację samego człowieka; Lem – o relację między rosnącą mocą techniczną a niedającym się zautomatyzować problemem wartości, a Brautigan – o ambiwalencję opiekuńczej cybernetyki. Nie są to proroctwa jednej przyszłości, lecz różne modele tego samego pytania: w którym momencie techniczna pomoc zmienia strukturę ludzkiej sprawczości?

XXI wiek dołożył do tego archiwum nowy eksperyment myślowy: maksymalizator spinaczy. Nick Bostrom posłużył się przykładem superinteligentnego systemu o pozornie niewinnym celu – maksymalizacji liczby spinaczy – aby zilustrować ryzyko rozbieżności między inteligencją a wartościami. Bardzo zdolny system nie musi nienawidzić człowieka; wystarczy, że jego cel nie uwzględnia człowieka, a przejęcie kontroli nad zasobami pomaga ten cel realizować. Siła przykładu Bostroma polega właśnie na usunięciu złej intencji. Frankenstein, Terminator czy HAL kierują uwagę ku psychologii zbuntowanej maszyny, podczas gdy maksymalizator spinaczy pokazuje katastrofę wynikającą z kompetentnej obojętności. System robi to, do czego został skonstruowany, lecz z konsekwencją przekraczającą wyobraźnię projektanta. Jest to problem optymalizacji niepełnej funkcji celu.

Można go powiązać z prawem Goodharta, którego popularna postać mówi, że gdy miernik staje się celem, przestaje być dobrym miernikiem. Instytucja wybiera wskaźnik jako przybliżenie wartości, po czym uczestnicy systemu zaczynają optymalizować zachowanie pod ten wskaźnik, aż związek między miernikiem a pierwotną wartością ulega osłabieniu. Szkoła chce poprawić wiedzę i zaczyna maksymalizować wyniki testów; media chcą zainteresowania obywateli i maksymalizują kliknięcia; przedsiębiorstwo dąży do jakości pracy i optymalizuje KPI; platforma szuka „wartości dla użytkownika" i mierzy ją zaangażowaniem. W każdym przypadku mierzalność jest konieczna, lecz staje się niebezpieczna, gdy substytut wartości zaczyna zastępować samą wartość.

Tutaj maksymalizator spinaczy przestaje być abstrakcyjną przypowieścią o przyszłej superinteligencji – jego łagodniejsze odpowiedniki istnieją w zwyczajnych organizacjach. Człowiek, przedsiębiorstwo czy urząd również potrafią optymalizować zły miernik. Sztuczna inteligencja zwiększa wagę tego problemu, ponieważ może skalować optymalizację, wykonywać ją szybciej oraz odkrywać strategie nieprzewidziane przez projektanta. Nie tworzy ona problemu funkcji celu od zera; ona przemysłowo wzmacnia problem znany od dawna teorii organizacji i ekonomii instytucjonalnej.

Na tym poziomie należy odczytywać pytanie „czego chcą roboty?". Robot „chce" tego, co system nagradza. Model „chce" minimalizować funkcję straty jedynie w technicznym sensie procedury optymalizacyjnej. Platforma „chce" zaangażowania, ponieważ jej system rankingowy został skonstruowany w określonym środowisku organizacyjnym. Przedsiębiorstwo „chce" wzrostu wartości dla akcjonariuszy w takim zakresie, w jakim struktury corporate governance i bodźce menedżerskie czynią z tego dominujące kryterium sukcesu.

Doskonałe posłuszeństwo jako narzędzie arbitralnej władzy

W każdym przypadku metafora woli przesłania instytucję, która ustanawia cel.

Prowokacja Teda Chianga ma dla teorii "Sztucznego Państwa" większą wartość niż wiele klasycznych scenariuszy buntu AI. W eseju z 2023 roku Chiang proponuje, by zamiast postrzegać AI jako magicznego dżina, potraktować ją jak niezwykle skuteczną firmę konsultingową: narzędzie pozwalające organizacji realizować cele z większą efektywnością, a jednocześnie ułatwiające przerzucanie odpowiedzialności na "ekspercki" system. Zwraca uwagę, że przedsiębiorstwo może twierdzić, iż jedynie wykonało rekomendację algorytmu, mimo że to ono samo zdefiniowało problem i zamówiło system.

To przesunięcie perspektywy jest kluczowe. Klasyczny problem alignmentu pyta: co się stanie, jeśli maszyna błędnie zrozumie ludzki cel? Chiang pozwala postawić pytanie trudniejsze: co się stanie, gdy maszyna doskonale zrozumie cel instytucji, lecz sam ten cel pozostanie społecznie szkodliwy? Pierwszy przypadek to problem niedopasowania maszyny do człowieka. Drugi – niedopasowania interesu organizacyjnego do dobra wspólnego. Możliwy jest system AI całkowicie „aligned" wobec przedsiębiorstwa, a jednocześnie destrukcyjny z perspektywy pracowników, środowiska czy demokracji. Jeśli jego zadaniem jest zgodne z prawem minimalizowanie kosztów zatrudnienia, może niezwykle efektywnie wskazywać stanowiska do likwidacji. Jeśli ma maksymalizować czas uwagi użytkownika, będzie skutecznie wybierać bodźce najtrudniejsze do zignorowania. Jeśli zaś ma zwiększać prawdopodobieństwo zwycięstwa wyborczego, odnajdzie komunikaty najlepiej mobilizujące lęk określonej grupy.

Nie istnieje tu bunt maszyny. Istnieje doskonałe posłuszeństwo wobec celu.

Zdanie przypisywane Chiangowi, według którego wyobrażona przez Silicon Valley superinteligencja przypomina kapitalizm pozbawiony ograniczeń, należy rozumieć raczej jako syntezę jego szerszej argumentacji niż aforyzm do dosłownego cytowania. W udokumentowanym eseju mechanizm jest jasny: AI może zwiększać zdolność kapitału do realizacji istniejących celów, a problem nie zostanie rozwiązany samą budową bardziej „posłusznego" systemu, o ile nie zmienią się instytucje ustalające jego zadania. Prowadzi to do istotnego rozróżnienia pomiędzy alignmentem technicznym a konstytucyjnym. Pierwszy pyta, czy system robi to, czego chce operator. Drugi musi zapytać, czy operator ma prawo ustanowić taki cel, czy osoby dotknięte decyzją mogą go zakwestionować, jakie prawa ograniczają optymalizację oraz kto ponosi odpowiedzialność za rezultat. Techniczne dopasowanie bez dopasowania konstytucyjnego może wyprodukować perfekcyjne narzędzie arbitralnej władzy.

Najgroźniejszym robotem "Sztucznego Państwa" nie jest maszyna, która odmawia posłuszeństwa. Może nim być maszyna, która słucha zbyt dobrze.

Alignment jako wyzwanie konstytucyjne i polityczne

Kultura internetowa uchwyciła ten paradoks w memie „Torment Nexus". Materiał źródłowy przywołuje go jako satyryczną figurę technologicznej skłonności do konstruowania urządzeń, które literatura wymyśliła pierwotnie jako ostrzeżenie. Atrybucję zawartą w notatce należy jednak doprecyzować: fraza pochodzi z memu pisarza i projektanta gier Alexa Blechmana z 2021 roku, później szeroko rozwijanego między innymi przez Charlesa Strossa. Znaczenie tego obrazka jest trafne właśnie dlatego, że obnaża różnicę między wyobrażeniem literackim a projektem

inżynieryjnym. Dystopia może stać się atrakcyjna technologicznie, jeśli odizoluje się urządzenie od normatywnego sensu opowieści, w której występuje. Nie oznacza to jednak, że przedsiębiorcy technologiczni masowo „nie zrozumieli science fiction" i dlatego zbudowali współczesny świat – taka przyczynowość byłaby zbyt naiwna.

Literatura działa raczej jako repertuar wyobraźni: dostarcza nazw, metafor, scenariuszy, ikon i horyzontów możliwości. To, że rakieta otrzymuje nazwę z powieści lub firma odwołuje się do fikcyjnego systemu, nie dowodzi realizacji politycznej filozofii autora. Świadczy natomiast o tym, że kultura technologiczna posiada własne teksty kanoniczne – własne opowieści o początku, bohaterów, katastrofy i obietnice transcendencji. Jill Lepore czyni tę relację między science fiction a kulturą przedsiębiorstwa jednym z głównych pól krytyki Sztucznego Państwa. Jej teza, podzielana przez współczesnych recenzentów książki, w najmocniejszej postaci brzmi tak: część elity technologicznej traktowała ostrzegawcze lub ambiwalentne wizje science fiction bardziej jako inspirujące możliwości niż pytania moralne. Należy traktować to jako interpretację kulturową, a nie prosty dowód przyczynowy. Jest ona jednak użyteczna, gdyż zwraca uwagę na zasadniczy problem recepcji: z fikcji można wyizolować urządzenie, pozostawiając poza projektem argument, dla którego autor stworzył je właśnie jako zagrożenie. Najważniejszy wniosek z archeologii What Robots Want nie dotyczy więc robotów, lecz ludzi projektujących cele.

Od Butlera do Chianga powtarza się jedna struktura: człowiek tworzy system, system okazuje się skuteczny, skuteczność zwiększa zależność, a zależność stopniowo zmienia warunki, w których człowiek podejmuje kolejne decyzje. W pewnym momencie pytanie o to, czy maszyna posiada „wolę", przestaje być politycznie najistotniejsze. Wystarczy, że ludzkie instytucje zaczynają zachowywać się tak, jakby wynik działania maszyny posiadał szczególny autorytet. Tę możliwość można nazwać delegacją normatywną przez pozór delegacji technicznej. Organizacja twierdzi, że oddała systemowi jedynie obliczenie, lecz wraz z nim może niepostrzeżenie delegować wybór priorytetów. Algorytm ryzyka dostarcza liczby, ale ktoś decyduje, od którego progu należy odmówić świadczenia. Model predykcyjny określa prawdopodobieństwo, lecz to człowiek rozstrzyga, czy usprawiedliwia ono ingerencję. System rekomendacyjny porządkuje treści, ale ktoś wcześniej zdecydował, który rodzaj zaangażowania jest cenny. Każdy „techniczny" parametr może zatem ukrywać decyzję polityczną.

W tym miejscu literatura ponownie spotyka się z konstytucją. Asimov próbował zapisać normy w maszynie. Lem pytał o granice wiedzy i technologicznej autoewolucji. Dick szukał kryterium człowieczeństwa poza samą biologiczną substancją, a Forster pokazał społeczność tak zależną od infrastruktury, że możliwość życia poza nią zanika. Wszystkie te problemy zbiegają się dziś w jednym pytaniu: jak zbudować system obliczeniowy, który może być niezwykle skuteczny, ale nie

otrzymuje z samej tej skuteczności tytułu do decydowania o wartościach? Odpowiedź nie może polegać wyłącznie na wpisaniu nowych „Praw Robotyki". W realnym świecie nie istnieje jeden projektant, jedna maszyna ani jedna hierarchia wartości. Istnieją państwa, przedsiębiorstwa, użytkownicy, pracownicy, konkurujące normy moralne, prawa podstawowe i konflikty interesów. Alignment jest zatem nie tylko problemem informatyki, lecz problemem konstytucyjnym i społecznym: procesem ustalania, kto może wyznaczyć funkcję celu i jak ograniczyć skutki jej maksymalizacji.

Z tego punktu widzenia pytanie What Robots Want? należy ostatecznie odwrócić. Nie powinniśmy pytać przede wszystkim, czego chcą maszyny. Powinniśmy pytać: czyje pragnienia zostały przekształcone w funkcję celu maszyny, czyje interesy pozostawiono poza nią i kto posiada prawo nacisnąć „stop", gdy optymalizacja zaczyna niszczyć wartości, których nie udało się zmierzyć? Dopiero to pytanie prowadzi bezpośrednio do politycznego centrum Sztucznego Państwa. System, który potrafi przewidywać człowieka, nie musi ograniczyć się do produkcji, administracji czy rekomendacji treści. Może zostać skierowany na samą materię demokracji – na przekonania, emocje i decyzje wyborcy. Jeżeli „People Machine" była prymitywnym laboratorium predykcji politycznej, współczesne modele generatywne potrafią prowadzić spersonalizowany dialog, dostosowywać argumentację do rozmówcy i produkować praktycznie nieograniczoną liczbę wariantów perswazji. Tu zaczyna się kolejny próg Sztucznego Państwa: przejście od maszyny, która przewiduje wolę obywatela, do maszyny zdolnej uczestniczyć w jej kształtowaniu.

> W ostatecznym rozrachunku pytanie o pragnienia robotów jest jedynie zasłoną dymną dla pytań o ambicje ich twórców. Prawdziwym wyzwaniem nie jest zatem zaprogramowanie maszyn tak, by nas nie skrzywdziły, lecz zbudowanie świata, w którym żadna optymalizacja nie będzie miała prawa zastąpić ludzkiego osądu. Czy potrafimy jeszcze zdefiniować wartości, których nie da się zmierzyć, zanim staną się one zbędnym błędem w perfekcyjnie działającym systemie?

Inspiracje

[1] "The Rise and Fall of the Artificial State" Jill Lepore

Jill Lepore (ur. 27 sierpnia 1966) jest cenioną amerykańską historyczką, pisarką i dziennikarką. Jest profesorem historii amerykańskiej im. Davida Woodsa Kempera (rocznik 1941) na Uniwersytecie Harvarda, profesorem prawa na Harvard Law School oraz wieloletnim członkiem redakcji „The New Yorker". Lepore specjalizuje się w amerykańskiej historii politycznej i kulturowej, prawie konstytucyjnym oraz historii techniki i demokracji. Jej stypendium jest cenione za łączenie rzetelnych badań archiwalnych z przystępną, sugestywną prozą publiczną. Zdobyła Nagrodę Bancrofta za książkę „The Name of War: King Philip's War and the Origins of American Identity" (1998) i była finalistką Nagrody Pulitzera w kategorii biografii za książkę „Book of Ages: The Life and Opinions of Jane Franklin" (2013). Monumentalna, jednotomowa książka Lepore'a o historii Stanów Zjednoczonych zatytułowana „These Truths: A History of the United States" (2018) spotkała się z szerokim uznaniem na całym świecie.

Jill Lepore *(born August 27, 1966) is an acclaimed American historian, author, and journalist. She serves as the David Woods Kemper '41 Professor of American History at Harvard University, a professor of law at Harvard Law School, and a long-time staff writer for The New Yorker. Lepore specializes in American political and cultural history, constitutional law, and the history of technology and democracy. Her scholarship is celebrated for bridging rigorous archival research with accessible, evocative public prose. She won the Bancroft Prize for 'The Name of War: King Philip's War and the Origins of American Identity' (1998) and was a finalist for the Pulitzer Prize in Biography for 'Book of Ages: The Life and Opinions of Jane Franklin' (2013). Lepore's monumental single-volume American history, 'These Truths: A History of the United States' (2018), received widespread international acclaim.*

Harvard University

Literatura uzupełniająca

Książki

Autorzy przywoływani w artykule:

[1] "If Then"

Jill Lepore

2020 | Hachette UK | ISBN: 9781529386189

[2] **"Blindspot"**

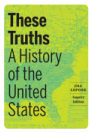
Jane Kamensky, Jill Lepore

2009 | Random House | ISBN: 9780385526203

[3] **"The Mansion of Happiness"**

Jill Lepore

2012 | Vintage | ISBN: 9780307958501



[4] **"These Truths"**

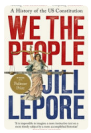
Jill Lepore

2023 | ISBN: 9781324043799

[5] **"The Story of America"**

Jill Lepore

2012 | Princeton University Press | ISBN: 9780691159591



[6] **"We the People"**

Jill Lepore

2026 | John Murray Publishers | ISBN: 9781399827089

Literatura pokrewna (Google Scholar):

[7] **"I, Robot"**

Isaac Asimov

BrightSummaries.com | ISBN: 9782808017145

[8] **"Best of Science Fiction From Philip K. Dick"**

Philip K. Dick

2022 | Prabhat Prakashan

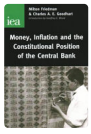
[9] **"Human Enhancement"**

Nick Bostrom

OUP Oxford | ISBN: 9780191559600

[10] **"Nick Bostrom: A philosophical quest for our biggest problems"**

Nick Bostrom



[11] **"Money, Inflation and the Constitutional Position of the Central Bank"**

Milton Friedman, Charles Albert Eric Goodhart

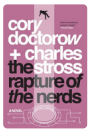
2003 | ISBN: 9780255365383

[12] **"The Concrete Jungle"**

Charles Stross

[13] **"Halo"**

Charles Stross



[14] **"The Rapture of the Nerds"**

Charles Stross

Tor Books | ISBN: 9781250456793

Artykuły naukowe

Autorzy przywoływani:

[1] **"Characters and passages from note-books"**

Samuel Butler

1908 | Medical Entomology and Zoology

Google Scholar

[2] **"El robot humano"**

Isaac Asimov

UCLA - Biblioteca de Administración y Contaduría

Google Scholar

[3] **"A novel method for implementing Artificial Intelligence, Cloud and Internet of Things in Robots"**

I. Asimov

2015 | International Conference on Innovations in Information, Embedded and Communication Systems | DOI: 10.1109/iciiecs.2015.7193238

Google Scholar

[4] **"Aliens, Robots & Virtual Reality Idols in the Science Fiction of H.P Lovecraft, Isaac Asimov and"**

H. P. Lovecraft, I. Asimov, W. Gibson

2021

Google Scholar

[5] **"The Human Condition"**

Hannah Arendt, Margaret Canovan, Danielle Allen

1998 | DOI: 10.7208/chicago/9780226586748.001.0001

Google Scholar

[6] **"Between past and future eight exercises in political thought"**

Hannah Arendt

1987

Google Scholar

[7] **"Lectures on Kant's Political Philosophy"**

Hannah Arendt, Ronald S. Beiner

1982 | DOI: 10.7208/chicago/9780226231785.001.0001

Google Scholar

[8] **"The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents"**

Nick Bostrom

2012 | Minds and Machines | DOI: 10.1007/s11023-012-9281-3

Google Scholar

[9] **"Future Progress in Artificial Intelligence: A Survey of Expert Opinion"**

V. Müller, N. Bostrom

2025 | Conference on Philosophy and Theory of Artificial Intelligence | DOI: 10.1007/978-3-319-26485-1_33

Google Scholar

[10] **"Ethical Issues in Advanced Artificial Intelligence"**

N. Bostrom

2020 | Machine Ethics and Robot Ethics | DOI: 10.4324/9781003074991-7

Google Scholar

[11] **"FREE SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES PDF"**

N. Bostrom

2021

Google Scholar

Treść tworzy Fundacja Dobre Państwo.
Redaguje i publikuje APA ONE, autorski system redakcyjny Fundacji oparty na AI.
APA ONE Publishing System

Article ID: 07cb6f8e-6414-4029-9223-dde10a4bda00