Aksjologia przyszłości: między optymalizacją a rozkwitem w świetle The Challenge of the Future



AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje przejście parametrów ludzkiego życia z domeny losu do domeny decyzji projektowych dzięki technologii i transhumanizmowi. Autor podkreśla kluczowe rozróżnienie między wartościami instrumentalnymi (środkami) a autotelicznymi i konstytutywnymi, które współtworzą dobre życie. Ostrzega przed redukcjonizmem ekonomicznym oraz ryzykiem błędnego zdefiniowania funkcji celu w systemach AI. Głównym zagrożeniem jest sytuacja, w której niezwykle sprawna technologia optymalizuje mierzalne wskaźniki (np. PKB czy produktywność), myląc je z rzeczywistym dobrostanem i ludzkim rozkwitem. W tym kontekście problem alignment w sztucznej inteligencji staje się nie tylko wyzwaniem technicznym, ale przede wszystkim głębokim pytaniem aksjologicznym o to, jakie wartości powinny być realizowane.

The text analyzes the transition of human life parameters from the domain of fate to the domain of design decisions through technology and transhumanism. The author emphasizes the key distinction between instrumental values (means) and autotelic and constitutive values, which together constitute a good life. He warns against economic reductionism and the risk of misdefining the objective function in AI systems. The main threat is a situation where highly efficient technology optimizes measurable indicators (e.g., GDP or productivity), confusing them with actual well-being and human flourishing. In this context, the problem of alignment in artificial intelligence becomes not only a technical challenge but, above all, a profound axiological question about which values should be realized.

W dobie gwałtownego rozwoju sztucznej inteligencji i biotechnologii, ludzkość staje przed bezprecedensowym wyzwaniem: przejście od życia determinowanego przez los do

życia będącego wynikiem projektowych decyzji. Artykuł stawia tezę, że największym zagrożeniem przyszłości nie jest sama technologia, lecz ryzyko błędu kategorialnego – pomylenia mierzalnych wskaźników (takich jak PKB, czas retencji w aplikacji czy subiektywna satysfakcja) z rzeczywistymi celami ludzkiego rozkwitu. Autor argumentuje, że optymalizacja techniczna bez nadbudowy w postaci mądrości praktycznej może prowadzić do dystopii, w której człowiek pozostaje zadowolony, tracąc jednocześnie podmiotowość i realne zdolności sprawcze (capability). Kluczowym postulatem jest odrzucenie redukcjonistycznego hedonizmu na rzecz pluralistycznej aksjologii, która rozróżnia dobra instrumentalne od konstytutywnych. W tym kontekście problem 'alignmentu' AI przestaje być jedynie zadaniem inżynieryjnym, a staje się fundamentalnym pytaniem cywilizacyjnym o to, jakimi istotami chcemy być i czy jesteśmy w stanie odróżnić inteligencję, która rozwiązuje problemy, od mądrości, która wie, które problemy są w ogóle warte rozwiązania.

SŁOWA KLUCZOWE

aksjologia przyszłości | alignment w sztucznej inteligencji | eudajmonia | capability approach | wartości autoteliczne | wartości instrumentalne | wartości konstytutywne | funkcja celu | optymalizacja technokratyczna | rozkwit człowieka | maszyna doświadczeń Nozicka | produkcja preferencji | podejście możliwości | mądrość praktyczna | transhumanizm

Technologia zmienia parametry losu w decyzje projektowe

Aksjologia przyszłości: co jest warte chcenia, gdy niemal wszystko staje się możliwe

Moment, w którym człowiek zaczyna projektować nie tylko swoje otoczenie, lecz także własne właściwości biologiczne i poznawcze, fundamentalnie zmienia status aksjologii. Przez znaczną część historii filozofia wartości była refleksją nad tym, jak godnie przeżyć życie w warunkach, których podstawowych parametrów nie wybieraliśmy: rodziliśmy się śmiertelni, podatni na choroby, ograniczeni możliwościami ciała i umysłu oraz zależni od natury, społeczności i zastanego porządku historycznego.

Technologia przesuwa jednak te parametry z domeny losu do domeny decyzji. Jeśli możemy modyfikować długość życia, płodność, zdolności poznawcze, wzorce zachowań czy genetyczne właściwości przyszłych pokoleń, pytanie o to, co jest dobre, przestaje być jedynie filozoficznym komentarzem do egzystencji.

Staje się częściowo specyfikacją projektową przyszłości. Właśnie dlatego w książce Graylinga refleksja nad sztuczną inteligencją i transhumanizmem prowadzi bezpośrednio do pojęcia wartości. Jako trzy drogi rozpoznawania tego, co istotne, wskazano obserwację życia społecznego, refleksję filozoficzną oraz literaturę utopijną; jednocześnie wprowadzono rozróżnienie między wartościami autotelicznymi a instrumentalnymi oraz krytykę redukcji dobrego życia do hedonistycznie rozumianego szczęścia. Ten fragment notatki wymaga rozwinięcia, gdyż dotyka problemu kluczowego dla oceny niemal każdej technologii przyszłości: czy potrafimy odróżnić to, co pomaga osiągnąć dobre życie, od tego, co samo w sobie stanowi jego element? Rozróżnienie to wydaje się banalne, dopóki nie zauważymy, jak często cywilizacje zamieniają środek w cel.

Pieniądz jest tu najbardziej oczywistym przykładem. Jego funkcja jest z natury instrumentalna: ułatwia wymianę i dostęp do dóbr. Jeśli jednak sukces społeczny definiujemy głównie przez akumulację kapitału, wartość instrumentalna zaczyna udawać autoteliczną.

Analogicznie technologia pozostaje narzędziem rozszerzania ludzkich możliwości. Sama moc obliczeniowa nie wskazuje kierunku jej użycia. Szybszy system może równie skutecznie diagnozować choroby, co rozpowszechniać dezinformację; lepszy model predykcyjny może poprawić bezpieczeństwo lub udoskonalić represje.

Instrument staje się dobry dopiero w relacji do celu, który realizuje. Współczesna teoria wartości rozróżnia zatem dobro cenione dla niego samego od dobra instrumentalnego, którego wartość wynika z umożliwiania osiągnięcia czegoś innego. Filozofia komplikuje jednak tę dychotomię poprzez pojęcie wartości konstytutywnej – nie jest ona kolejnym narzędziem prowadzącym do dobra, lecz jego integralną częścią. Przyjaźń nie musi być środkiem do dobrego życia w taki sposób, w jaki samochód jest środkiem transportu; może sama współtworzyć życie uznawane za dobre. Podobnie wiedza: może służyć zdobyciu zawodu, ale samo poznawanie może stanowić istotę ludzkiego rozkwitu.

Stanford Encyclopedia of Philosophy zauważa, że wartość instrumentalna odnosi się typowo do relacji przyczynowej środek-cel, podczas gdy wartość konstytutywna opisuje to, co dobro współtworzy, zamiast jedynie je powodować. To rozróżnienie pozwala doprecyzować ujęcie, w którym edukacja została podana jako przykład wartości instrumentalnej służącej „dobremu życiu”. Jest to interpretacja możliwa, lecz zbyt wąska z perspektywy pluralistycznej teorii wartości.

Edukacja może być instrumentem prowadzącym do zatrudnienia czy mobilności społecznej, ale rozwój zdolności rozumienia świata, argumentacji czy poznanie kultury mogą być konstytutywnymi składnikami rozwoju osoby. Kto traktuje edukację wyłącznie jako urządzenie zwiększające produktywność kapitału ludzkiego, dokonuje redukcji charakterystycznej dla społeczeństwa ekonomicznego: interpretuje wartość człowieka przez pryzmat produktywności, a wartość wiedzy przez jej rynkową stopę zwrotu.

Ryzyko pomylenia mierzalnych wskaźników z celem życia

To właśnie takie błędy kategorialne mogą okazać się jednym z największych zagrożeń przyszłości. Jeśli pomylimy środki z celami, niezwykle sprawna technologia nie skoryguje pomyłki, lecz ją zwielokrotni. System optymalizacyjny, który doskonale realizuje źle określoną funkcję celu, jest groźniejszy od systemu nieefektywnego. Ta obserwacja łączy aksjologię z problemem alignment w sztucznej inteligencji: nie wystarczy sprawić, aby system skutecznie realizował zadanie; trzeba jeszcze odpowiedzieć na pytanie, czy właściwie określiliśmy wartość, której realizacji oczekujemy.

Ten sam problem przejawia się na poziomie społecznym. Państwo maksymalizujące PKB może jednocześnie pogarszać zdrowie psychiczne obywateli, niszczyć środowisko, podrywać zaufanie społeczne lub psuć jakość relacji – o ile wzrost gospodarczy zostanie błędnie potraktowany jako dobro ostateczne, a nie jedno z narzędzi umożliwiających realizację szerszego zestawu dóbr. Aksjologia przyszłości musi zatem rozpocząć się od podejrzliwości wobec wszystkich pojedynczych wskaźników dobra. Nie dlatego, że pomiar sam w sobie jest zły, lecz dlatego, że każda metryka redukuje wielowymiarową rzeczywistość do jednego, wybranego aspektu. Produkt krajowy brutto mierzy produkcję rynkową, nie znaczenie życia; liczba przeżytych lat mierzy długość, nie jakość biografii; liczba interakcji w sieci mierzy aktywność, nie głębokość relacji; liczba publikacji może mierzyć produktywność naukową, lecz niekoniecznie poznawczą doniosłość.

Kiedy metryka staje się celem systemu, ludzie i instytucje zaczynają adaptować swoje zachowania do sposobu pomiaru. W konsekwencji instrument obserwacji może przeobrazić obserwowany świat. Problem ten nabiera szczególnej ostrości w środowisku algorytmicznym. System rekomendacyjny potrzebuje funkcji celu: czasu spędzonego na platformie, kliknięć, zakupów, retencji, satysfakcji deklarowanej przez użytkownika lub innego mierzalnego rezultatu. Żaden z tych wskaźników nie jest jednak tożsamy z ludzkim dobrem.

Człowiek może spędzać więcej czasu na platformie dlatego, że otrzymuje wartościowe treści, ale również dlatego, że system skutecznie wzbudza w nim niepokój, oburzenie albo kompulsywną ciekawość. Wzrost zaangażowania jest zatem faktem behawioralnym, a nie oceną moralną. Algorytm może idealnie maksymalizować wskaźnik, którego związek z ludzkim rozkwitem jest słaby lub wręcz ujemny. Aksjologia nie jest więc luksusem dołączanym do techniki po zakończeniu prac konstrukcyjnych; pojawia się już w chwili wyboru tego, co ma być maksymalizowane. Najbardziej intuicyjnym kandydatem na dobro ostateczne wydaje się szczęście.

Problem polega jednak na wieloznaczności samego słowa. W literaturze przedmiotu rozróżnia się krótkotrwały stan emocjonalny – occurrent sentiment – od trwałych okoliczności życia, takich jak wolność, zdrowie, edukacja i bezpieczeństwo, które tworzą infrastrukturę dobrostanu. To rozróżnienie jest analitycznie bardzo produktywne, ponieważ wpisuje się w znacznie szerszy spór filozoficzny. Współczesna filozofia dobrostanu zwykle wyróżnia trzy wielkie rodziny teorii: hedonizm, teorie spełnienia pragnień oraz teorie listy obiektywnej. Każda z nich inaczej odpowiada na pytanie, co właściwie sprawia, że życie idzie dobrze dla osoby, która je przeżywa. Hedonizm prudencjalny odpowiada najprościej: dobre dla człowieka jest to, co zwiększa korzystny bilans przyjemności i cierpienia.

Siła tej teorii wynika z bezpośredniości doświadczenia. Ból wydaje się zły nie dlatego, że społeczeństwo umówiło się go potępiać, lecz dlatego, że jest przeżywany jako cierpienie. Podobnie przyjemność posiada oczywistą wartość fenomenologiczną. Nie należy więc lekceważyć hedonizmu jako prymitywnej filozofii konsumpcji.

Eliminowanie bólu, głodu, lęku i cierpienia stanowi jedno z największych osiągnięć medycyny, polityki społecznej i cywilizacji. Problem zaczyna się wtedy, gdy hedonizm z prawdziwej tezy „cierpienie ma znaczenie” przechodzi do znacznie mocniejszej tezy: „tylko bilans przyjemności i cierpienia ma znaczenie”.

Klasyczny eksperyment myślowy Roberta Nozicka z „maszyną doświadczeń” uderza dokładnie w tę intuicję. Gdyby można było podłączyć człowieka do urządzenia zapewniającego mu przez resztę życia subiektywne doświadczenie sukcesu, miłości, przyjaźni i twórczości, wiele osób intuicyjnie odmówiłoby wejścia do maszyny. Chciałyby nie tylko doświadczać przekonania, że kochają i są kochane, lecz rzeczywiście kochać i być kochane; nie tylko czuć satysfakcję z napisania książki, lecz faktycznie ją napisać. Współczesna filozofia dobrostanu używa właśnie takich argumentów, aby wskazać, że samo doświadczenie może nie wyczerpywać tego, co jest dla nas dobre.

Eksperyment Nozicka zyskuje nową aktualność w świecie technologii zdolnych dostarczać coraz bardziej spersonalizowanych środowisk doświadczenia. Wirtualna rzeczywistość, algorytmicznie

komponowana rozrywka, syntetyczni towarzysze czy potencjalne neurotechnologie regulujące emocje pozwalają postawić to pytanie znacznie poważniej niż w XX-wiecznej filozoficznej fikcji. Jeśli można będzie technologicznie zwiększać subiektywną satysfakcję bez proporcjonalnej poprawy obiektywnych okoliczności życia, polityka stanie wobec wyboru między poprawą rzeczywistości a poprawą odczuwania rzeczywistości. Najbardziej dystopijna wersja systemu nie musi zatem zadawać ludziom bólu. Może skutecznie sprawiać, że pozostają zadowoleni z położenia, którego w warunkach pełnej wiedzy i autonomii nie zaakceptowaliby.

Tu właśnie „Nowy wspaniały świat" Huxleya jest być może bardziej subtelną dystopią niż Orwellowski świat jawnego terroru.

Deklarowana satysfakcja nie jest dowodem wolności

Władza nie musi karać, jeśli potrafi regulować pragnienia. System nie musi odbierać człowiekowi wszystkiego, czego ten chce; może raczej wcześniej ukształtować to, czego człowiek będzie chciał. Problem etyczny przesuwa się wówczas z dystrybucji dóbr na produkcję preferencji. Jednostka deklaruje satysfakcję, lecz ta przestaje być oczywistym dowodem wolności.

Odpowiedzią na ograniczenia hedonizmu stały się między innymi teorie spełnienia pragnień. Według nich życie człowieka jest dobre, jeśli realizowane są jego rzeczywiste cele i preferencje. Koncepcja ta trafnie oddaje różnicę między subiektywnym doświadczeniem sukcesu a sukcesem rzeczywistym: ktoś może pragnąć faktycznie ukończyć badanie, wychować dziecko lub zbudować trwałą przyjaźń, zamiast jedynie doświadczać psychologicznej symulacji tych osiągnięć. Współczesna teoria dobrostanu wskazuje jednak na słabość tego stanowiska: pragnienia mogą być błędnie poinformowane, chwilowe, autodestrukcyjne, wymuszone presją lub dostosowane do skrajnie złych warunków. To ostatnie zagadnienie ma kluczowe znaczenie społeczne.

Osoba funkcjonująca przez lata w warunkach silnych ograniczeń może dostosować swoje aspiracje do tego, co jest osiągalne. Jeśli miarą sprawiedliwości uczynimy wyłącznie deklarowane zadowolenie lub aktualne preferencje, możemy uznać sytuację za dobrą tylko dlatego, że jednostka nauczyła się nie oczekiwać niczego więcej. Ekonomia dobrobytu, socjologia nierówności i teoria polityczna muszą zatem rozróżniać preferencję od możliwości wyboru. Adaptacja do braku wolności nie przekształca braku wolności w dobro.

Problem komplikuje się jeszcze bardziej w środowisku cyfrowym, gdzie preferencje przestają być jedynie danymi wejściowymi systemu, a stają się jego rezultatem. Algorytm rekomendacyjny obserwuje wcześniejsze wybory i na ich podstawie przewiduje kolejne; prezentuje treści

zwiększające prawdopodobieństwo określonych reakcji, by następnie wykorzystać te reakcje – wywołane przecież przez poprzednie rekomendacje – jako nowe dane.

Wartości obiektywne i eudajmonia przeciwko dyktatowi danych

Preferencja staje się w tym układzie endogeniczna: system nie tylko dowiaduje się, czego człowiek chce, lecz częściowo współtworzy przestrzeń, w której ten zaczyna czegoś pragnąć. Klasyczna liberalna formuła „dajmy ludziom to, czego chcą" okazuje się epistemicznie niepełna, gdy przemysł dysponuje narzędziami do kształtowania tych pragnień. Teorie listy obiektywnej próbują uniknąć obu trudności, zakładając, że pewne składniki życia są dobre dla człowieka niezależnie od tego, czy w danej chwili wywołują przyjemność lub są przez niego pożądane. Typowe przykłady obejmują wiedzę, przyjaźń, autonomię, osiągnięcia, zdrowie czy sensowne relacje. Współczesna teoria dobrostanu wskazuje właśnie wiedzę i przyjaźń jako dobra, których wartości nie należy redukować do poziomu przyjemności czy spełnienia doraźnych zachcianek. Atutem takiego podejścia jest możliwość krytyki społeczeństwa, którego mieszkańcy deklarują zadowolenie mimo braku podstawowych warunków rozwoju. Ceną jest jednak konieczność odpowiedzi na trudne pytanie: kto ustala tę listę i skąd wie, które dobra rzeczywiście mają status obiektywny? Spór ten nie daje się rozstrzygnąć prostym plebiscytem.

Jeżeli zapytamy ludzi, co cenią, otrzymamy ważną informację o ludzkim doświadczeniu, lecz sama popularność danej wartości nie dowodzi jej normatywnej prawdziwości. Społeczeństwa przez stulecia aprobowały instytucje, które dziś oceniamy jako głęboko niesprawiedliwe. Z drugiej strony filozof nie może po prostu sporządzić własnego katalogu dobra i ogłosić go obowiązującym dla wszystkich. Dlatego Grayling zestawia trzy źródła poznania wartości: potoczne doświadczenie, refleksję filozoficzną i literaturę jako laboratorium możliwych światów. Metodę tę można wzbogacić o nauki empiryczne, historię instytucji, psychologię, antropologię czy ekonomię, jednak żadne z tych pól nie zastąpi namysłu normatywnego.

Wartości nie da się bowiem po prostu „wydobyć z danych". Z miliardów cyfrowych śladów możemy wnioskować, na czym ludzie spędzają czas, co kupują i czego szukają, ale nie wynika z tego, że właśnie te zachowania ujawniają to, co powinni uważać za dobro. Dane behawioralne opisują wybory dokonane w określonej architekturze bodźców. Człowiek może spędzić godzinę na kompulsywnym przewijaniu treści, jednocześnie uznając, że lepsza byłaby godzina rozmowy z bliską osobą.

System mierzy zachowanie; aksjologia pyta, czy było ono zgodne z bardziej refleksyjną koncepcją dobra. W tym kontekście szczególną wagę odzyskuje arystotelesowska eudaimonia. Termin ten często tłumaczy się jako „szczęście", lecz współczesne interpretacje filozoficzne trafniej oddają go przez „flourishing" – rozkwit lub dobre funkcjonowanie całego życia. Etyka Arystotelesa rozpoczyna się od pytania o eudajmonię i przechodzi do analizy aretē, czyli doskonałości lub cnoty oraz cech charakteru niezbędnych, by człowiek mógł żyć dobrze. Grayling wykorzystuje tę tradycję przeciw redukowaniu najwyższego dobra do chwilowego stanu hedonicznego, wiążąc eudajmonię z mądrym i cnotliwym życiem. Kluczowa różnica polega na tym, że eudajmonia jest kategorią biograficzną, a nie chwilową.

Człowiek może doświadczać bólu, żałoby czy frustracji, a mimo to uważać życie za głęboko wartościowe. Rodzicielstwo, badania naukowe, twórczość, działalność publiczna czy opieka nad chorym mogą wiązać się z ogromnymi trudnościami, których nie da się przedstawić jako stałego strumienia przyjemności. Ich znaczenie wynika z zaangażowania w dobra i relacje przekraczające chwilowy afekt. Dobre życie nie jest więc koniecznie życiem, w którym w każdej chwili „dobrze się czujemy".

Ta intuicja stawia poważne pytanie technologicznej kulturze wygody. Jeżeli systemy cyfrowe coraz skuteczniej redukują wysiłek, oczekiwanie i niepewność, nie oznacza to, że proporcjonalnie zwiększają rozkwit. Wysiłek posiada wartość instrumentalną tylko wtedy, gdy prowadzi do czegoś sensownego; nie ma powodu, by romantyzować ból czy niedostatek. Jednak niektóre zdolności kształtują się właśnie poprzez praktykę, odroczoną gratyfikację i konieczność podejmowania decyzji bez pełnej pewności. Całkowite usunięcie oporu może zatem nie tylko ułatwiać życie, ale zmieniać proces formowania osoby. Pytanie o „zbędne cnoty", obecne w materiale, ma więc głębokie znaczenie aksjologiczne. Sugeruje ono, że w świecie całkowitej automatyzacji odwaga, cierpliwość czy mądrość mogą utracić środowisko, w którym były dotychczas praktykowane. Hipotezy tej nie należy jednak traktować jako oczywistej prawdy – możliwe, że technologia nie unicestwi cnót, lecz zmieni ich przedmiot.

Wartości jako realne możliwości działania

Odwaga może dziś mniej dotyczyć fizycznej walki, a bardziej obrony niepopularnej prawdy; umiarkowanie – nie niedostatku jedzenia, lecz radzenia sobie z nadmiarem bodźców; cierpliwość – nie oczekiwania na usługę, lecz trwania w długotrwałym procesie poznawczym; mądrość zaś nie magazynowania informacji, lecz rozpoznawania, kiedy jest ona prawdziwa, relewantna i moralnie

znacząca. Taka perspektywa chroni humanizm przed konserwatywnym błędem utożsamiania wartości z historyczną formą ich realizacji.

Odwaga nie wymaga wojny, podobnie jak solidarność nie potrzebuje katastrofy. Zadaniem postępu nie powinno być zatem konserwowanie cierpienia tylko po to, by ludzie mieli okazję wykazywać się cnotą. Powinno nim być raczej usuwanie cierpienia zbędnego przy jednoczesnym zachowaniu przestrzeni dla odpowiedzialnego działania, nauki i relacji. Grayling wyraźnie rozróżnia tu cierpienie naturalne od moralnego, kładąc silny akcent na kategorię bólu możliwego do uniknięcia. Aksjologia przyszłości nie może więc bronić bólu dlatego, że ból historycznie uczestniczył w kształtowaniu charakteru.

Bardziej obiecującym pomostem między eudajmonią a współczesną teorią polityczną okazuje się capability approach Amartyi Sena i Marthy Nussbaum. Jego centralna intuicja polega na przesunięciu uwagi z posiadanych zasobów na rzeczywiste możliwości działania i bycia. Dwie osoby o identycznym dochodzie nie muszą bowiem dysponować takimi samymi realnymi szansami – różnią się zdrowiem, środowiskiem społecznym, doświadczeniem dyskryminacji czy dostępnością infrastruktury.

Capability approach traktuje wolność osiągania dobrostanu jako wartość zasadniczą. Analizuje dobro nie tylko przez pryzmat osiągniętych rezultatów (funkcjonowań), lecz przede wszystkim przez realny zbiór opcji, spośród których człowiek może wybierać. To rozróżnienie ma kluczowe znaczenie dla naszej technologicznej przyszłości. Jeśli AI przejmuje coraz więcej zadań, możemy obserwować paradoks: jednoczesny wzrost rezultatów przy spadku capability. Uczeń otrzymujący doskonały tekst z systemu posiada produkt, ale niekoniecznie zdolność samodzielnego skonstruowania argumentu. Pracownik wspomagany automatyzacją zwiększa produktywność, lecz może tracić wiedzę niezbędną do działania bez wsparcia systemu. Osoba polegająca na algorytmicznym doradcy może podejmować statystycznie lepsze decyzje, a jednocześnie słabiej rozwijać zdolność samodzielnego osądu.

W perspektywie capability nie wystarczy zatem pytać, czy technologia poprawiła wynik. Należy zapytać, czy zwiększyła realną zdolność człowieka do działania jako podmiotu. W tym punkcie capability approach koryguje zarówno naiwny technoprogresywizm, jak i bezkrytyczny luddystyczny opór. Nie każda delegacja zadań oznacza bowiem utratę autonomii.

Wózek elektryczny zwiększa capability osoby z ograniczeniami mobilności właśnie dlatego, że urządzenie wykonuje pracę, której człowiek sam wykonać nie może. Tłumacz automatyczny może realnie rozszerzać możliwość uczestniczenia w edukacji i życiu społecznym, a system wspierający diagnostykę – zwiększać zdolność lekarza do skutecznego wykrywania choroby.

Pluralizm wartości przeciwko technokratycznej optymalizacji

Kluczowe jest pytanie o relację między narzędziem a sprawstwem: czy technologia poszerza zakres realnych możliwości człowieka, czy też zastępuje kompetencje w sposób prowadzący do zależności od infrastruktury kontrolowanej przez inny podmiot. Z tej perspektywy autonomia staje się wartością bardziej złożoną niż sama możliwość formalnego wyboru. Człowiek może mieć prawo do odrzucenia algorytmicznego systemu, ale jeśli bez niego nie jest w stanie uzyskać pracy, kredytu, ubezpieczenia czy edukacji, a nawet uczestniczyć w podstawowych praktykach społecznych, jego wybór staje się iluzoryczny. Capability approach podkreśla właśnie różnicę między prawem formalnym a rzeczywistą możliwością korzystania z niego. W epoce AI problem wolności będzie więc coraz częściej sprowadzał się nie do pytania o to, co jest prawnie dozwolone, lecz o to, czy jednostka posiada materialne, poznawcze i instytucjonalne zasoby, by skutecznie dokonać wyboru.

Tak rozumiana wolność prowadzi ku pluralizmowi wartości. Trudno znaleźć mocne podstawy dla założenia, że wszystkie dobra można bez reszty przeliczyć na jedną jednostkę. Miłość, wiedza, autonomia, bezpieczeństwo, twórczość, zdrowie, sprawiedliwość i przyjaźń mogą być wartościami częściowo niewspółmiernymi. W teorii wartości spór między monizmem a pluralizmem dotyczy właśnie tego, czy wszystkie ostateczne dobra można sprowadzić do jednej podstawowej wartości, czy też istnieje ich wiele. Pluralizm komplikuje politykę, ale lepiej oddaje doświadczenie ludzkiego życia. Część prawdziwych tragedii moralnych wynika bowiem nie z konfliktu dobra ze złem, lecz ze starcia dwóch rzeczywiście cennych dóbr.

Wolność może kolidować z bezpieczeństwem, prywatność z efektywnością diagnostyczną, równość z pewnymi formami swobody ekonomicznej, ochrona przyszłych pokoleń z bieżącą konsumpcją, a autonomia pacjenta z paternalistycznie rozumianym interesem zdrowotnym. Problemów tych nie da się zawsze rozwiązać poprzez zwiększenie ilości jednego dobra. Potrzebne są procedury ważenia, kompromisu i wyznaczania granic, które ochronią pewne wartości przed całkowitym poświęceniem dla innych. Aksjologia przyszłości nie będzie więc fabryką jednej funkcji celu dla całej ludzkości. Powinna raczej tworzyć język pozwalający rozpoznawać konflikty wartości bez maskowania ich jako problemów technicznych. To jest zarazem miejsce, w którym do teorii wartości wchodzi sprawiedliwość. Jeżeli technologia zwiększa łączny dobrobyt, ale jej zyski przejmuje niewielka grupa, nie wystarczy odpowiedzieć, że suma korzyści wzrosła. Trzeba zapytać o dystrybucję, możliwości najsłabszych, równość pozycji i zasady uzasadniania nierówności. Grayling wyraźnie włącza do katalogu wartości sprawiedliwość i wolność, przeciwstawiając im dyskryminację algorytmiczną i ucisk, a bezpieczeństwu ekonomicznemu – ubóstwo i wykluczenie. Wartość

społeczna nie daje się więc sprowadzić do agregatu indywidualnych satysfakcji. Przykład sztucznej inteligencji dobrze obrazuje tę różnicę.

Załóżmy, że automatyzacja generuje ogromny wzrost produktywności, powodując jednocześnie koncentrację własności i utratę wielu stabilnych źródeł dochodu. Z perspektywy czysto agregacyjnej system może wyglądać korzystnie, gdyż całkowita produkcja wzrosła. Z perspektywy sprawiedliwości pytamy jednak: kto posiada technologię, kto czerpie rentę z produktywności, kto ponosi koszt transformacji i czy osoby zastąpione przez system zachowują realne możliwości uczestnictwa społecznego. Właśnie takie napięcie pomiędzy obietnicą uwolnienia od pracy a ryzykiem powstania klasy ekonomicznie zbędnej pojawia się w książce Graylinga. Nie jest więc wystarczające twierdzenie, że technologia „zwiększa dobrobyt". Trzeba za każdym razem zapytać: czyj dobrobyt, w jakim znaczeniu, mierzony czym, w jakim okresie i kosztem kogo? To pięć prostych pytań rozbija znaczną część technokratycznej retoryki.

System może zwiększać dobrobyt średni, pogarszając sytuację osób na dole rozkładu; poprawiać dobrostan krótkookresowy kosztem długookresowego; zwiększać satysfakcję konsumenta kosztem prywatności obywatela; obniżać ceny usług kosztem erozji warunków pracy; zapewniać wygodę obecnemu pokoleniu kosztem zasobów przyszłych. Do aksjologii należy zatem również czas.

Postęp techniczny nie gwarantuje postępu moralnego

Wartość przyszłego człowieka nie maleje z powodu tego, że jeszcze się nie urodził. Choć ekonomia posługuje się dyskontowaniem przyszłych przepływów pieniężnych, przeniesienie tej logiki bezpośrednio na wartość osoby jest filozoficznie problematyczne. Jeśli szkoda klimatyczna nastąpi za pięćdziesiąt lat, późniejszy termin nie czyni cierpienia przyszłego człowieka metafizycznie mniej realnym. Technologia o wielopokoleniowych konsekwencjach wymaga zatem aksjologii intertemporalnej – teorii odpowiedzialności uwzględniającej osoby nieobecne przy dzisiejszym stole negocjacyjnym.

W tym miejscu przyszłość staje się problemem demokracji. Dzisiejszy wyborca może nagradzać politykę przynoszącą korzyści w ciągu najbliższych kilku lat, korporacja odpowiada przed inwestorami oceniającymi wyniki kwartalne, a polityk działa w krótkim cyklu wyborczym. Tymczasem klimat, infrastruktura, genom, składowanie odpadów, system emerytalny czy przestrzeń orbitalna generują skutki rozłożone na dziesięciolecia. Instytucje o krótkim horyzoncie decyzyjnym zarządzają więc procesami o bardzo długim horyzoncie oddziaływania, co tworzy zjawisko, jakim jest strukturalna asymetria czasu. Aksjologia przyszłości musi ponadto uwzględniać fakt, że wartości ewoluują w czasie.

W tym kontekście warto przywołać koncepcję Jaspersowskiego „Przełomu Osiowego", rozumianą jako przejście od etyki siły i cnót wojownika ku refleksyjnej moralności oraz cnotom obywatelskim. Ujęcie to służy jako cenna heurystyka, lecz nie może być traktowane jako dowód linearnego postępu moralnego. Jaspers dostrzegał w pierwszym tysiącleciu p.n.e. równoległe przemiany intelektualne w Chinach, Indiach, Iranie, Izraelu i Grecji, związane ze wzrostem refleksyjności etycznej i krytyką zastanego porządku. Niemniej sama koncepcja „Axial Age" pozostaje przedmiotem sporów; badacze wskazują zarówno na różnice między tymi tradycjami, jak i na elementy moralności rozwijające się znacznie wcześniej. Najważniejsza lekcja z historii wartości nie polega zatem na przekonaniu, że ludzkość nieuchronnie zmierza ku doskonałości. Historia nie jest ruchomymi schodami prowadzącymi ku dobru. Można mówić o wzroście w konkretnych dziedzinach – rozszerzaniu katalogu praw, potępieniu niewolnictwa czy ochronie przed arbitralnością władzy – a jednocześnie obserwować regres, nowe formy przemocy i techniczne udoskonalanie starych patologii. Współczesna filozofia postępu odróżnia teleologiczną ideę historii zmierzającej koniecznie do celu od ujęcia otwartego, w którym postęp zależy od działań ludzi i instytucji, a przez to może zostać zatrzymany lub odwrócony. To rozróżnienie współgra z główną tezą całego artykułu.

Nie istnieje bowiem żadne historyczne prawo gwarantujące, że większa inteligencja techniczna wygeneruje lepszą moralność. Społeczeństwo może stać się technologicznie bardziej złożone, a jednocześnie normatywnie bardziej brutalne. Narzędzia Oświecenia mogą zostać obrócone przeciw jego ideałom; zaawansowana administracja może zwiększać prawa obywatela albo doskonalić kontrolę; AI może poszerzać dostęp do wiedzy lub indywidualizować propagandę. Postęp technologiczny i postęp moralny to odrębne zmienne. Właśnie dlatego pojęcie „alignment" powinno zostać wyjęte z wąskiej debaty o sztucznej inteligencji i potraktowane jako problem cywilizacyjny. Zazwyczaj pytamy, jak dopasować cele systemów AI do wartości ludzkich. Ale o jakie wartości chodzi? O te deklarowane przez obecnych użytkowników? Preferencje konsumentów? Prawa człowieka? Wartości większości, projektantów czy prawodawstwa poszczególnych państw? A może o interes przyszłych pokoleń?

Alignment AI jako lustro ludzkich sprzeczności

Odpowiedzialny alignment wymaga zatem nie tylko rozwiązania inżynieryjnego, lecz solidnej teorii politycznej i aksjologicznej. Wyzwaniem nie jest samo nauczenie maszyny naszych wartości, lecz fakt, że my sami nie jesteśmy w pełni zgodni co do ich hierarchii, a nasze praktyki często pozostają w sprzeczności z tym, co deklarujemy.

Próba zbudowania AI „idealnie zgodnej z człowiekiem” może w ten sposób ujawnić niewygodną prawdę o nas samych. Społeczeństwo głosi równość, akceptując jednocześnie rażące nierówności szans; deklaruje troskę o przyszłe pokolenia, kontynuując działania obarczone długoterminowymi kosztami środowiskowymi; chroni prywatność w zapisach prawnych, a równocześnie masowo wymienia dane osobiste na wygodę; uznaje autonomię za wartość, lecz projektuje systemy gospodarcze oparte na coraz dokładniejszym przewidywaniu i modyfikowaniu zachowań. Alignment systemu technicznego z taką kulturą może więc utrwalić jej sprzeczności, zamiast pomóc w ich rozwiązaniu.

Potrzebujemy czegoś więcej niż "chmury słów" wartości. Społeczeństwo musi wypracować mechanizmy refleksyjnego filtrowania preferencji – umiejętność odróżniania impulsu od przekonania, interesu od zasady, wartości deklarowanej od praktykowanej oraz środka od celu. Właśnie taką rolę może pełnić filozofia: nie jako urząd wydający ostateczny katalog dobra, lecz jako praktyka ujawniania sprzeczności i stawiania pytań, których system optymalizacyjny sam z siebie nigdy nie zada.

Warto w tym miejscu przywołać prosty zestaw wartości obecny w materiale: miłość i rodzina, praca i osiągnięcia, sprawiedliwość i wolność, bezpieczeństwo ekonomiczne, zdrowie i sprawność. Katalog ten nie musi pretendować do miana uniwersalnej teorii dobra, by pełnić funkcję heurystyczną. Jego siłą jest właśnie heterogeniczność – łączy relacje, aktywność, instytucje, zasoby i kondycję biologiczną, pokazując, że dobre życie nie składa się z jednego tworzywa.

Każda z tych wartości posiada również swój cień. Miłość zakłada możliwość straty, autonomia – możliwość błędu, osiągnięcie – ryzyko porażki, sprawiedliwość – konflikt roszczeń, a zdrowie świadomość cielesnej kruchości. Nie oznacza to, że należy pielęgnować negatywy, lecz że wiele dóbr ma strukturę relacyjną i dynamiczną. Świat zaprojektowany tak, by całkowicie wyeliminować rozczarowanie, mógłby jednocześnie usunąć przestrzeń dla autonomicznego działania. Świat bez ryzyka nie musi być bezpieczny dla podmiotowości, jeśli ceną tego bezpieczeństwa jest całkowita kontrola. Pojawia się zatem pytanie o wartość wolności samej w sobie.

Mądrość jako niezbędna nadbudowa nad inteligencją

Gdyby algorytm potrafił zawsze dokonywać za człowieka wyborów prowadzących do większego dobrostanu, czy należałoby oddać mu stery? Paternalistyczna odpowiedź brzmiałaby: tak, skoro wynik jest lepszy. Perspektywa liberalna odpowiada: nie, ponieważ możliwość autorstwa własnej biografii posiada wartość niezależną od jakości poszczególnych rezultatów. Capability approach wzmacnia tę drugą intuicję, uznając realną wolność osiągania dobrostanu za wartość moralnie

centralną. To rozróżnienie między dobrym wynikiem a samym autorstwem wyniku stanie się jednym z kluczowych problemów ery sztucznej inteligencji.

Można wyobrazić sobie system, który bezbłędnie dobiera partnera, zawód, miejsce zamieszkania, dietę czy strategię finansową. Statystycznie człowiek osiągałby lepsze rezultaty, lecz jego biografia przestałaby być projektem, który tworzy samodzielnie. Czy życie dobrze zoptymalizowane jest automatycznie życiem dobrze przeżytym? Tradycja eudajmonistyczna skłania do odpowiedzi przeczącej: rozkwit wymaga nie tylko osiągnięcia określonych stanów, lecz aktywności podmiotu dysponującego praktycznym rozumem. Arystotelesowska mądrość praktyczna nie polega na mechanicznym stosowaniu uniwersalnej instrukcji; wymaga doświadczenia, deliberacji i zdolności rozpoznawania tego, co właściwe w konkretnej sytuacji. Jeśli tak, to delegowanie decyzji posiada granicę aksjologiczną niezależną od jej efektywności. Nie chodzi tu o zakaz korzystania z doradców czy systemów eksperckich – człowiek od zawsze czerpał z cudzej wiedzy.

Problem pojawia się w momencie, gdy z procesu znika zdolność rozumienia, kwestionowania i ostatecznego przyjęcia odpowiedzialności. Podmiotowość nie wymaga, aby człowiek samodzielnie obliczał każdy parametr. Wymaga jednak, by nie utracił pozycji tego, dla kogo rachunek jest wykonywany i kto potrafi zakwestionować przyjętą funkcję celu. Dochodzimy tu do fundamentalnego rozróżnienia całej aksjologii przyszłości: inteligencja nie jest tym samym co mądrość. Inteligencję można rozumieć jako zdolność rozwiązywania problemów, rozpoznawania wzorców, przewidywania i optymalizacji. Mądrość wymaga dodatkowo rozpoznania, które problemy są w ogóle warte rozwiązania, których celów nie należy realizować, jakie koszty są moralnie niedopuszczalne i kiedy możliwość techniczna powinna ustąpić przed ograniczeniem normatywnym.

Inteligencja odpowiada na pytanie „jak?”. Mądrość zaczyna się od pytania „po co?” i nie kończy, dopóki nie zapyta „czyim kosztem?”. Im bardziej sztuczna inteligencja przejmuje funkcje pierwszego rodzaju, tym większego znaczenia nabierają te drugie. Paradoksalnie rozwój AI może zatem nie zmniejszać potrzeby filozofii, lecz ją potęgować.

Jeżeli maszyna potrafi znaleźć tysiąc sposobów realizacji celu, człowiek musi precyzyjniej wiedzieć, dlaczego dany cel został wybrany. Jeśli technologia potrafi wydłużyć życie, musimy lepiej rozumieć, czego oczekujemy od tych dodatkowych lat. Jeśli biotechnologia pozwala zmieniać cechy przyszłego człowieka, musimy wiedzieć, które różnice uznajemy za chorobę, a które za zwykłą różnorodność i gdzie „ulepszanie” byłoby jedynie projekcją aktualnych uprzedzeń na przyszłe pokolenia.

Mądrość w wyborze celów zamiast technokratycznej optymalizacji

Aksjologia nie dostarczy w tych sprawach gotowej listy poleceń. Jej funkcją jest raczej zapobieganie najgroźniejszemu złudzeniu technokratycznemu: przekonaniu, że problem zostaje rozwiązany w momencie, gdy zostanie wystarczająco dokładnie policzony. Nie każdy konflikt wartości wynika z braku danych. Czasem dysponujemy pełnią wiedzy i mimo to musimy wybierać między dobrami. To właśnie w tym punkcie zaczynają się polityka, etyka i odpowiedzialność.

Przyszłe społeczeństwo może zatem dysponować niewyobrażalnym bogactwem środków, pozostając jednocześnie ubogie w cele. Może opanować sztukę maksymalizacji życia, wydajności, przyjemności czy uwagi, nie wiedząc przy tym, dlaczego uznało właśnie te wielkości za miarę sukcesu. To aksjologiczny odpowiednik problemu źle sformułowanej funkcji celu: system działa bez zarzutu, lecz perfekcja wykonania jedynie szybciej oddala go od tego, czego naprawdę potrzebował. Dlatego kluczowe pytanie o przyszłość nie brzmi „ile więcej będziemy mogli mieć?", lecz „jakimi istotami chcemy być i jakie warunki umożliwiają nam stawanie się nimi?".

Współczesna refleksja przeciwstawia kulturę posiadania wartościom zakorzenionym w zdrowiu, refleksji i wewnętrznym ładzie osoby. Krytykę konsumpcjonizmu nie należy jednak sprowadzać do moralizowania przeciw materii. Zasoby są niezbędne dla wolności, bezpieczeństwa i uczestnictwa w życiu społecznym. Błąd pojawia się dopiero wtedy, gdy to, co ma jedynie umożliwiać życie, zaczyna zastępować odpowiedź na pytanie o to, czym życie powinno być. Przyszłość nie zostanie zatem rozstrzygnięta przez wybór między technologią a wartościami – taki dylemat jest fałszywy, gdyż każda technologia już teraz operuje wewnątrz określonej hierarchii wartości. Prawdziwy spór dotyczy tego, które z nich zostaną wpisane w funkcje celu, prawo, standardy dostępu i codzienne praktyki.

Możemy skonstruować AI jako narzędzie zwiększające ludzkie sprawstwo albo jako mechanizm pogłębiający zależność. Możemy wykorzystać biotechnologię do ograniczania cierpienia lub do reprodukcji biologicznych hierarchii. Możemy zwiększyć produktywność i zamienić jej zysk w czas wolny, albo skoncentrować rentę, czyniąc część populacji ekonomicznie zbędną.

Technologia poszerza zakres ruchu, lecz nie wybiera kierunku. Pierwszym zadaniem aksjologii przyszłości nie jest zatem wymyślanie nowych wartości, lecz uporządkowanie tych starych: rozpoznanie, co traktujemy jako cele, a co jako środki; co stanowi fundament dobrego życia, a co granice, których nie wolno przekroczyć nawet w imię korzyści agregatowej. Dopiero wtedy można sensownie pytać o kierunki rozwoju technicznego. Jednak nawet najstaranniej uporządkowana

hierarchia wartości napotyka kolejny problem: od wieków projektujemy miejsca, w których sprawiedliwość, bezpieczeństwo i szczęście miały zostać osiągnięte jednocześnie. Nazywaliśmy je dobrym państwem, idealną republiką czy utopią. Historia tych eksperymentów pokazuje jednak niepokojącą regularność: niemal każda wizja idealnego świata ukrywa koszt, którego nie widać na fasadzie.

Dlatego po aksjologii musi nadejść analiza utopii. Skoro dobra są wielorakie i często niewspółmierne, trzeba zapytać, co dzieje się wtedy, gdy ktoś próbuje zaprojektować system usuwający te napięcia raz na zawsze. To właśnie tam – pomiędzy Tomaszem Morusem, Nową Atlantydą Bacona, utopiami emancypacyjnymi, Huxleyem, Orwellem i Omelas Ursuli K. Le Guin – ujawnia się jedna z najważniejszych prawidłowości: każda obietnica doskonałego porządku powinna zostać przeczytana razem z drobnym drukiem dotyczącym jego ceny.

> Każda obietnica technologicznego raju posiada swój 'drobny druk' z ukrytymi kosztami, których nie dostrzeże żaden algorytm optymalizacyjny. Prawdziwym pytaniem nie jest to, czy potrafimy zbudować system idealnie zgodny z naszymi pragnieniami, lecz czy w świecie pozbawionym oporu i ryzyka nie utracimy tego, co czyni nas ludźmi. Czy życie perfekcyjnie zoptymalizowane może być jednocześnie życiem dobrze przeżytym?

Inspiracje

[1] "The Challenge of the Future" A C Grayling

Anthony Clifford Grayling (ur. 3 kwietnia 1949 r.) jest brytyjskim filozofem, pisarzem i akademikiem. Jest założycielem i magistrem Northeastern University London (pierwotnie założonego jako New College of the Humanities) oraz supernumerary fellow St Anne's College w Oksfordzie. Wcześniej był profesorem filozofii na Birkbeck, University of London, gdzie wykładał w latach 1991–2011. Główne zainteresowania naukowe Graylinga obejmują epistemologię, metafizykę, logikę, etykę i historię idei. Jako wybitny intelektualista i orędownik humanizmu, wniósł istotny wkład w debaty na temat etyki świeckiej, instytucji demokratycznych i swobód obywatelskich. W ponad trzydziestu książkach Grayling dąży do szerokiego udostępnienia badań filozoficznych, podejmując jednocześnie palące wyzwania społeczne i polityczne współczesnego świata.

Anthony Clifford Grayling (born April 3, 1949) is a British philosopher, author, and academic. He is the founder and Master of Northeastern University London (originally founded as the New College of the Humanities) and a supernumerary fellow of St Anne's College, Oxford. Prior to this, he served as Professor of Philosophy at Birkbeck, University of London, where he taught from 1991 to 2011. Grayling's core academic interests span epistemology, metaphysics, logic, ethics, and the history of ideas. A prominent public intellectual and advocate of humanism, he has contributed extensively to debates on secular ethics, democratic institutions, and civil liberties. Across more than thirty books, Grayling has worked to make philosophical inquiry broadly accessible while addressing pressing social and political challenges of the contemporary world.

Northeastern University London

Literatura uzupełniająca

Książki

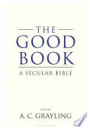
Książki autora i dzieła przywołane w tekście:



[1] "Grace"

A C Grayling

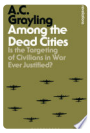
Oberon Books



[2] "The Good Book"

A C Grayling

A&C Black | ISBN: 9780747599609



[3] "Among the Dead Cities"

A C Grayling

A&C Black | ISBN: 9781472534057



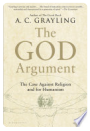
[4] "The Meaning of Things"

A C Grayling

Weidenfeld & Nicolson | ISBN: 9781780221168

[5] ""Bomber County": A riveting story of the air war"

A C Grayling

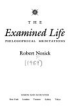


[6] "The God Argument"

A C Grayling

Bloomsbury+ORM | ISBN: 9781620401910

Książki autorów przywoływanych i literatura pokrewna:



[7] "The Examined Life"

Robert Nozick

Simon & Schuster

[8] "Dialogue of Comfort against Tribulation"

Tomasz Morus

ISBN: 9781312905054

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Development as Freedom"

Amartya Kumar Sen

2009 | DOI: 10.1007/978-1-137-28787-8_94

Źródło

[2] **"Well-Being, Agency and Freedom: The Dewey Lectures 1984"**
Amartya Kumar Sen
1985 | The Journal of Philosophy | DOI: 10.2307/2026184
Źródło

[3] **"Collective Choice and Social Welfare"**
H. Myint, Amartya Kumar Sen
1972 | Econometrica | DOI: 10.2307/1913866
Źródło

[4] **"Human Rights and Capabilities"**
Amartya Kumar Sen
2005 | Journal of Human Development | DOI: 10.1080/14649880500120491
Źródło

[5] **"Women and Human Development"**
Martha C. Nussbaum
2000 | Cambridge University Press eBooks | DOI: 10.1017/cbo9780511841286
Źródło

[6] **"Human Functioning and Social Justice"**
Martha C. Nussbaum
1992 | Political Theory | DOI: 10.1177/0090591792020002002
Źródło

[7] **"The challenge of gender justice"**
Martha C. Nussbaum
2009 | Against Injustice | DOI: 10.1017/cbo9780511657443.006
Źródło

Treść tworzy Fundacja Dobre Państwo.
Redaguje i publikuje APA ONE, autorski system redakcyjny Fundacji oparty na AI.
APA ONE Publishing System
Fundacja Dobre Państwo © 2026
Article ID: fc497aee-aa9d-4a03-9f19-b49c214ffd55