Architektura odpowiedzialności i kontroli w systemach hybrydowych człowieka i AI w kontekście The AI Instinct Rany Gujral

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Tekst analizuje problem odpowiedzialności w systemach hybrydowych człowiek-AI, przeciwstawiając kontrolę nominalną (formalną obecność człowieka) kontroli rzeczywistej (meaningful human control). Autor wskazuje, że etyka takich układów zależy od ich architektury, a nie tylko od moralności poszczególnych elementów. Kluczowe znaczenie mają koncepcje trackingu (responsywność systemu na ludzkie racje moralne i kontekst) oraz Racingu (możliwość przypisania działania do konkretnego podmiotu). Podejście to ma na celu rozwiązanie problemu luki odpowiedzialności (responsibility gap), gdzie złożoność techniczna rozprasza wiedzę i sprawstwo, czyniąc z człowieka jedynie „administracyjną" przeszkodę w procesie decyzyjnym, zamiast realnego nadzorcy.

The text analyzes the problem of responsibility in human-AI hybrid systems, contrasting nominal control (the formal presence of a human) with actual control (meaningful human control). The author indicates that the ethics of such systems depend on their architecture, rather than solely on the morality of individual components. Key importance is placed on the concepts of tracking (the system's responsiveness to human moral reasons and context) and tracing (the ability to attribute an action to a specific agent). This approach aims to resolve the problem of the responsibility gap, where technical complexity disperses knowledge and agency, rendering the human merely an "administrative" obstacle in the decision-making process rather than a real supervisor.

Artykuł podejmuje problematykę etycznej integracji człowieka ze sztuczną inteligencją, stawiając tezę, że tradycyjny nadzór nad technologią – często sprowadzany do

formalnego zatwierdzania decyzji przez operatora – jest niewystarczający i tworzy niebezpieczną iluzję kontroli. Autor argumentuje, że w obliczu postępującej hybrydyzacji poznawczej (koncepcja Fused Mind) musimy przejść od nadzoru nominalnego do architektury 'kontroli rzeczywistej' (meaningful human control). Oznacza to odejście od szukania pojedynczego winnego na rzecz systemowego projektu odpowiedzialności, który obejmuje m.in. responsywność systemu wobec ludzkich racji moralnych, transparentną strukturę delegowania uprawnień oraz mechanizmy aktywnego oporu i odwracalności decyzji. Tekst dowodzi, że suwerenność człowieka w świecie AI nie polega już na poznawczej samowystarczalności, lecz na zdolności do konstytucyjnego zarządzania własnymi zależnościami od technologii, co ma chronić naszą integralność psychiczną przed manipulacją i systemowym dryfem.

SŁOWA KLUCZOWE

meaningful human control | systemy hybrydowe człowiek-AI | architektura odpowiedzialności | responsibility gap | moral crumple zone | graduated autonomy | proaktywne tarcie | cognitive circuit breaker | contestability | integralność poznawcza | epistemiczna uczciwość | AI Act nadzór ludzki | tracking i tracing | alignment of incentives

Kontrola rzeczywista zamiast nominalnej obecności człowieka

Etyka sprzężenia człowiek-AI: meaningful human control, odwracalność, niepewność i odpowiedzialność w systemach hybrydowych

W momencie, gdy człowiek i sztuczna inteligencja zaczynają tworzyć jeden funkcjonalny system poznawczy, tradycyjna etyka technologii staje przed problemem bardziej złożonym niż ocena moralna pojedynczego narzędzia. Nie wystarczy zapytać o dokładność modelu, dobrą wiarę jego twórcy czy formalną zgodę użytkownika na jego zastosowanie. Etycznie istotny rezultat powstaje bowiem wewnątrz sprzężenia obejmującego projektanta, dane, model i interfejs, a także organizację wdrażającą, osobę podejmującą decyzję, podmiot podlegający jej konsekwencjom oraz przepisy określające granice dopuszczalnego działania. Moralna jakość hybrydy jest więc własnością jej architektury odpowiedzialności, a nie wyłącznie moralności któregokolwiek z jej elementów.

W systemie, w którym AI proponuje rozwiązanie, człowiek je zatwierdza, organizacja narzuca presję czasu, interfejs eksponuje jedną opcję jako domyślną, a producent modelu nie ujawnia wszystkich ograniczeń – stwierdzenie, że "ostateczna decyzja należała do człowieka", opisuje zaledwie ostatni moment długiego procesu przyczynowego. Dlatego etyka hybrydowej kognicji musi wyjść od pojęcia kontroli rzeczywistej, a nie nominalnej.

Człowiek może być formalnie obecny w procesie, naciskać przycisk zatwierdzający i widnieć w dokumentacji jako osoba odpowiedzialna, a jednocześnie mieć znikomą zdolność modyfikowania przebiegu zdarzeń. Może nie znać źródeł rekomendacji, nie mieć czasu na analizę, brakować mu kompetencji do niezależnej weryfikacji wyniku lub pozostawać pod presją organizacji, by rekomendacji nie kwestionować. W takim układzie human in the loop staje się kategorią administracyjną, nie etyczną.

Znacznie bardziej wymagająca koncepcja meaningful human control pyta nie o fizyczną obecność człowieka, lecz o to, czy system rzeczywiście pozostaje responsywny wobec ludzkich racji oraz czy istnieje podmiot zdolny zrozumieć jego możliwości, konsekwencje i miejsce w łańcuchu odpowiedzialności. Filippo Santoni de Sio i Jeroen van den Hoven zaproponowali jedną z najbardziej wpływowych rekonstrukcji tego pojęcia, odwołując się do filozofii odpowiedzialności i koncepcji guidance control.

Wyróżnili dwa podstawowe warunki: tracking i tracing. Warunek tracking oznacza, że działanie systemu powinno pozostawać responsywne wobec istotnych ludzkich racji moralnych oraz właściwości konkretnej sytuacji. Warunek tracing wymaga natomiast możliwości powiązania zachowania systemu z co najmniej jednym człowiekiem posiadającym dostateczne rozumienie techniczne i moralne jego funkcjonowania.

Rozróżnienie to pokazuje, że kontrola nie sprowadza się do możliwości zatrzymania maszyny. System może posiadać przycisk awaryjny, a jednocześnie działać w sposób niedostatecznie związany z wartościami i intencjami osób, które mają go kontrolować. Warunek tracking należy odczytywać szerzej niż jako literalną zgodność z poleceniem; system może dokładnie wykonywać instrukcję i zarazem pozostawać niewrażliwy na racje, które dla człowieka uzasadniają odstępstwo od niej. W praktyce społecznej reguły są bowiem często stosowane kontekstowo.

Lekarz może zmienić standardową procedurę ze względu na stan pacjenta, sędzia uwzględnia nietypowy konflikt zasad, pracownik administracji dostrzega sytuację wyjątkową, a kierowca odstępuje od reguły ruchu, by umożliwić przejazd pojazdowi ratunkowemu. System kontrolowany w sensie meaningful human control powinien reagować nie tylko na formalny zapis reguły, lecz również na istotne dla jej stosowania okoliczności. Z tego powodu kontrola człowieka nie polega

wyłącznie na możliwości wydania komendy, ale na takim skonstruowaniu systemu, aby jego działanie pozostawało reason-responsive.

Warunek tracing prowadzi z kolei do problemu rozproszonej odpowiedzialności. Im bardziej system jest złożony, tym trudniej znaleźć osobę posiadającą pełne rozumienie wszystkich jego warstw: producent modelu nie zna kontekstu wdrożenia, integrator zna aplikację, lecz nie historię treningu, organizacja zna procedurę, ale nie techniczne własności systemu, a operator zna konkretny przypadek, lecz może nie rozumieć mechanizmów modelu. W rezultacie wiedza konieczna do przypisania odpowiedzialności zostaje rozproszona.

Tracing nie powinien zatem oznaczać poszukiwania jednego "winnego człowieka", lecz zaprojektowanie łańcucha odpowiedzialności, w którym jasno określono, kto odpowiada za model, integrację, dane, procedurę, uprawnienia i konkretny akt użycia. Jest to bezpośrednia odpowiedź na problem responsibility gap. W systemach autonomicznych może dojść do sytuacji, w której skutek jest moralnie i prawnie doniosły, ale żaden pojedynczy człowiek nie posiadał dostatecznej kontroli, by klasyczne przypisanie odpowiedzialności było przekonujące. Nie wolno tego rozwiązywać poprzez sztuczne cofnięcie odpowiedzialności do najbliższego operatora. Prowadziłoby to do powstania moral crumple zone, w której człowiek absorbuje konsekwencje awarii całego systemu mimo ograniczonego rzeczywistego wpływu na jego działanie.

Etycznie dojrzała automatyzacja wymaga zatem, aby wzrost autonomii maszyny był równoważony nie rytualną obecnością człowieka, lecz precyzyjniejszą konstrukcją instytucjonalnej odpowiedzialności. Unijne prawo sztucznej inteligencji odzwierciedla podobną intuicję w odniesieniu do systemów wysokiego ryzyka. Artykuł 14 AI Act wymaga projektowania takich systemów w sposób umożliwiający skuteczny nadzór osób fizycznych, przy czym środki nadzoru mają być proporcjonalne do ryzyka, stopnia autonomii i kontekstu zastosowania.

Skuteczna kontrola jako architektura stopniowanej autonomii

Kluczowe jest tutaj znaczenie przymiotnika „skuteczny”. Nadzór, który istnieje jedynie technicznie, lecz nie pozwala na realną zmianę działania systemu, nie realizuje funkcji, dla której został wprowadzony. Prawo dociera w tym miejscu do punktu zbieżnego z kognitywistyką: kontrola wymaga zdolności wpływania na proces, a nie tylko obserwowania jego wyniku.

Proponowany protokół bezpieczeństwa obejmuje mechanizmy takie jak pytanie przed działaniem w decyzjach wysokiej stawki, prezentowanie dowodów i alternatyw, ujawnianie niepewności,

możliwość wycofania decyzji oraz proaktywne tarcie. Elementy te nie powinny być traktowane jako zwykła checklista interfejsu, lecz jako fundament konstytucji systemu hybrydowego. Każdy z nich odpowiada bowiem innemu rodzajowi zagrożenia: autoryzacja chroni przed nieuprawnioną autonomią wykonawczą, alternatywy przeciwdziałają monopolowi poznawczemu, ujawnianie niepewności chroni przed pozorną pewnością, rollback ogranicza nieodwracalność, natomiast tarcie poznawcze ma zapobiegać automatyzacji ludzkiej zgody.

Mechanizm „ask before acting" należy interpretować przez pryzmat stopnia ryzyka i odwracalności, a nie jako uniwersalny nakaz. System poprawiający literówkę nie powinien żądać osobnej zgody na każdą zmianę. Natomiast agent przelewający znaczne środki, wysyłający wiążące oświadczenie, modyfikujący dane medyczne, podejmujący działanie prawne lub uruchamiający system fizyczny, powinien napotkać znacznie wyższy próg samodzielności. Istotą rozsądnego nadzoru nie jest zatem maksymalizacja liczby kliknięć „zatwierdź", lecz prawidłowa identyfikacja punktów nieodwracalności i wysokiej stawki. Można tu przywołać analogię do systemów bezpieczeństwa procesowego: w elektrowniach, lotnictwie czy medycynie nie każda decyzja wymaga tego samego poziomu autoryzacji.

Warstwy ochronne projektuje się proporcjonalnie do potencjalnej szkody. Podobnie agentowa AI powinna posiadać ściśle określone granice uprawnień. Może ona autonomicznie wykonywać działania odwracalne i niskiego ryzyka, wymagać potwierdzenia przy działaniach średniej stawki oraz całkowicie tracić możliwość automatycznego wykonania w przypadkach, gdy prawo, etyka lub skala możliwej szkody wymagają osobistej decyzji podmiotu uprawnionego. Jest to architektura delegacji, a nie prosty podział na człowieka i maszynę.

Taka konstrukcja prowadzi do pojęcia graduated autonomy. Autonomia systemu nie musi być cechą binarną; może zależeć od domeny, kwoty, rodzaju danych, liczby osób dotkniętych skutkiem, odwracalności, poziomu niepewności oraz historii niezawodności systemu. Agent może mieć pełną swobodę w organizacji dokumentów, ograniczoną w negocjowaniu terminów i całkowity brak swobody w składaniu wiążących oświadczeń finansowych. Co więcej, zakres autonomii powinien być dynamiczny – system wykazujący serię błędów może automatycznie przechodzić w bardziej restrykcyjny tryb nadzoru. Właśnie tutaj AGE w słabym funkcjonalnym sensie spotyka się z etyką: historia błędów powinna modyfikować przyszłe uprawnienia systemu.

Drugim filarem jest prezentowanie alternatyw. Przyjmuje się, że system nie powinien ograniczać użytkownika do jednej odpowiedzi, lecz przedstawiać uzasadnienie i przynajmniej jedną opcję alternatywną. Intuicję tę należy jednak rozbudować.

Epistemiczna uczciwość i odwracalność jako warunki realnej kontroli

Generowanie kilku wariantów przez ten sam model nie gwarantuje epistemicznego pluralizmu. Alternatywy mogą wydawać się powierzchownie różne, podczas gdy w rzeczywistości opierają się na tej samej ukrytej reprezentacji problemu. W decyzjach o wysokiej stawce znacznie bardziej wartościowe jest przedstawienie konkurencyjnych modeli problemu: innej hipotezy przyczynowej, odmiennego założenia normatywnego, źródła danych przeczącego rekomendacji lub scenariusza, w którym proponowana decyzja okazuje się błędna.

W tym sensie system powinien pełnić nie tylko funkcję doradcy, ale również red teamu własnego wyniku. Nie chodzi tu o generowanie sztucznej opozycji dla każdej propozycji, lecz o instytucjonalizację epistemicznej krytyki. W tradycyjnej nauce rolę tę pełnią recenzenci, konkurencyjne laboratoria i reprodukcja wyników; w sądownictwie strony procesu; w demokracji opozycja i pluralizm mediów; w przedsiębiorstwie zaś audyt oraz mechanizmy kontroli wewnętrznej. System hybrydowy powinien posiadać odpowiednik takiego kontradyktoryjnego mechanizmu wszędzie tam, gdzie jedna generatywna odpowiedź mogłaby zbyt łatwo stać się poznawczym monopolem.

Trzeci filar – ujawnianie niepewności – jest zadaniem znacznie trudniejszym, niż sugeruje prosty wskaźnik procentowy. Niepewność może bowiem występować na różnych poziomach. System może być niepewny co do faktów z powodu braku danych; może działać poza rozkładem sytuacji znanych z treningu; może napotykać sprzeczne źródła lub posiadać rzetelną wiedzę o faktach, lecz mierzyć się z nierozstrzygniętym konfliktem wartości. Może wreszcie być niepewny z powodu samych ograniczeń modelu. Każda z tych sytuacji wymaga odmiennego sposobu komunikacji.

Podanie jednej liczby, np. „pewność 82%”, może wręcz tworzyć fałszywe poczucie matematycznej precyzji. Z punktu widzenia metapoznania ważniejsza od surowej pewności jest kalibracja. System dobrze skalibrowany powinien mylić się częściej wtedy, gdy wskazuje niższą pewność, i rzadziej w sytuacjach, gdy deklaruje wysoką. Dla człowieka współpracującego z AI jest to kluczowe, ponieważ zaufanie można racjonalnie dostosować tylko wtedy, gdy sygnały niepewności mają przewidywalny związek z realną zawodnością. System stale deklarujący wysoką pewność mimo znacznego poziomu błędów generuje nie tylko problem techniczny, lecz etyczny: jego styl komunikacji zniekształca zdolność użytkownika do podjęcia świadomej decyzji.

Wysokiej jakości komunikacja niepewności powinna odróżniać brak wiedzy od konfliktu wiedzy. Jeśli system nie posiada dostępu do aktualnych danych, powinien to wprost zakomunikować. Jeśli

istnieją dwa poważne stanowiska naukowe, powinien przedstawić ten konflikt, zamiast wygładzać go do jednej syntetycznej odpowiedzi. W przypadku problemów o charakterze normatywnym nie może symulować istnienia jednej statystycznie poprawnej odpowiedzi. Tym samym ujawnianie niepewności jest nie tylko funkcją statystyczną, ale formą epistemicznej uczciwości systemu.

Czwarty element – rollback – wprowadza do etyki sztucznej inteligencji kategorię odwracalności. Gujral postuluje możliwość natychmiastowego cofnięcia skutków decyzji algorytmicznej, jednak w praktyce pełna odwracalność nie zawsze jest możliwa. Wysłanej wiadomości nie da się „odczytać"; decyzji medycznej nie zawsze można cofnąć; ujawnionych danych nie da się „odzobaczyć", a działanie fizyczne może prowadzić do trwałej szkody. Etyka systemów agentowych powinna zatem odróżniać odwracalność techniczną, proceduralną i kompensacyjną. Odwracalność techniczna oznacza możliwość cofnięcia operacji do stanu poprzedniego. Proceduralna polega na istnieniu ścieżki wzruszenia decyzji lub przywrócenia wcześniejszego statusu prawnego. Kompensacyjna natomiast odnosi się do możliwości naprawienia szkody, gdy cofnięcie samego zdarzenia jest niemożliwe.

Im słabsza jest odwracalność, tym wyższy powinien być próg autonomicznego działania AI. Można zatem sformułować ogólną zasadę: dopuszczalna autonomia systemu powinna maleć wraz ze wzrostem nieodwracalności i skali potencjalnej szkody. Zasada ta ma szczególne znaczenie dla agentów działających w środowisku cyfrowym.

Projektowanie mechanizmów odwracalności i aktywnego oporu

Niektóre działania wydają się łatwo odwracalne, lecz w rzeczywistości wywołują wtórne konsekwencje. Agent może usunąć plik, który da się odzyskać, ale w międzyczasie uruchomić proces zależny od jego braku. Może zmienić konfigurację, którą później cofniemy, jednak zmiana ta zdąży już wpłynąć na zachowanie innych systemów. Może opublikować informację, której techniczne usunięcie nie cofa faktu jej rozpowszechnienia. Dlatego rollback należy rozumieć jako projektowanie ścieżek odzyskiwania kontroli, a nie naiwną wiarę, że wszystkie skutki automatyzacji można anulować za pomocą narzędzi technicznych.

Piątym elementem jest proaktywne tarcie. Gujral definiuje je jako celowe spowalnianie systemu, aby przed przyjęciem sugestii AI aktywować bardziej refleksyjny tryb myślenia. Po wcześniejszej analizie procesów dualnych należy tę intuicję sformułować precyzyjniej: nie chodzi o dosłowne „włączenie Systemu 2", jak gdyby człowiek posiadał przełącznik między dwoma modułami. Chodzi

raczej o zmianę struktury sytuacji tak, aby zwiększyć prawdopodobieństwo niezależnego przetwarzania informacji, zauważenia konfliktu, rekonstrukcji przesłanek lub świadomego zaakceptowania ryzyka. Tarcie może przyjmować różne formy.

System może wymagać samodzielnego wstępnego osądu przed ujawnieniem rekomendacji, poprosić o potwierdzenie krytycznych założeń, wskazać najbardziej prawdopodobny scenariusz błędu, przedstawić kontrargument lub wymusić dodatkową autoryzację przy zmianie parametrów działania. Zasadą nie powinno być jednak mnożenie przeszkód; nadmiar ostrzeżeń prowadzi do habituacji i zjawiska alert fatigue. Człowiek przyzwyczajony do kilkudziesięciu komunikatów bezpieczeństwa zaczyna je ignorować. Właściwie zaprojektowane tarcie musi więc być rzadkie, informacyjne i proporcjonalne do wagi decyzji. Można powiedzieć, że tarcie jest odpowiednikiem hamulca poznawczego. Hamulec w samochodzie nie jest zaprzeczeniem mobilności, lecz warunkiem bezpiecznej prędkości. Podobnie mechanizmy opóźniające określone decyzje nie są przeciwieństwem efektywności AI – umożliwiają one bezpieczne korzystanie z głębszej automatyzacji tam, gdzie bez nich szybkość przekształciłaby się w niekontrolowany dryf. Paradoksalnie to właśnie projektowanie skutecznych mechanizmów zatrzymania pozwala systemowi działać bardziej autonomicznie w pozostałych etapach procesu.

Z tego wynika pojęcie cognitive circuit breaker. W finansach automatyczne przerwanie handlu ma zapobiegać samowzmacniającej się dynamice rynku w sytuacjach gwałtownych zmian. Analogiczny mechanizm w systemach poznawczych mógłby uruchamiać dodatkowy nadzór, gdy AI wykryje konflikt źródeł, nietypową sytuację, gwałtowny spadek pewności, niespójność między agentami, odejście od dotychczasowej polityki lub szczególnie wysoką potencjalną szkodę. Celem nie jest udowodnienie, że system zawsze potrafi rozpoznać własny błąd, lecz wykorzystanie mierzalnych sygnałów ryzyka do automatycznej zmiany trybu kontroli. Takie rozwiązanie musi jednak zachować niezależność od procesu, który nadzoruje. System nie może być jedynym arbitrem określającym, kiedy sam wymaga kontroli – przypominałoby to organizację, której jedynym audytorem jest jednostka poddawana audytowi. Potrzebne są mechanizmy heterogeniczne: drugi model, deterministyczna reguła, niezależny moduł bezpieczeństwa, ograniczenie uprawnień, zewnętrzny monitoring lub człowiek. Kluczowym pojęciem staje się defence in depth – wielowarstwowość bezpieczeństwa, w której awaria jednego elementu nie eliminuje wszystkich mechanizmów kontroli.

Etyka sprzężenia musi również obejmować contestability, czyli możliwość zakwestionowania rezultatu systemu. Jest to wartość szersza niż transparentność. Można doskonale wyjaśnić, dlaczego algorytm podjął daną decyzję, a jednocześnie nie dać osobie żadnego skutecznego sposobu jej zmiany. Contestability wymaga instytucjonalnej procedury: prawa do wiedzy, że automatyzacja

odegrała istotną rolę, możliwości przedstawienia informacji pominiętej przez system, dostępu do człowieka zdolnego dokonać rzeczywistej rewizji oraz mechanizmu naprawienia błędnego rozstrzygnięcia. Jest to szczególnie ważne w administracji, finansach, zatrudnieniu, medycynie i wymiarze sprawiedliwości, gdzie decyzja nie jest zwykłą rekomendacją konsumencką, lecz wpływa na prawa i możliwości jednostki. Contestability pełni również funkcję epistemiczną dla samego systemu: odwołania i sprzeciwy dostarczają informacji o przypadkach, których automatyzacja nie obsłużyła prawidłowo. Instytucja pozbawiona procedury odwoławczej nie tylko krzywdzi jednostkę, ale traci dostęp do danych o własnej zawodności. Dobrze skonstruowany system odwoławczy jest więc kanałem społecznego uczenia. W perspektywie AGE można stwierdzić, że contestability staje się jednym z mechanizmów, dzięki którym instytucja hybrydowa transformuje się po własnych błędach. Prowadzi to do pojęcia organizacyjnej rekonsolidacji pamięci.

Jeżeli błąd AI zostaje naprawiony jedynie w pojedynczym przypadku, system nie zdobywa doświadczenia. Jeżeli natomiast incydent prowadzi do aktualizacji procedury, zmiany reguł eskalacji, poprawy danych, nowego szkolenia ludzi i modyfikacji modelu, przeszłe zdarzenie staje się częścią przyszłej architektury instytucji.

Odpowiedzialność systemowa zamiast szukania kozła ofiarnego

Jest to odpowiednik słabego AGE (Artificial General Intelligence) na poziomie organizacyjnym. Organizacja zmienia się, ponieważ wcześniej zawiodła. Etyka wymaga w takim przypadku uczenia się na incydentach, lecz bez kultury kozła ofiarnego. Jeśli każdy błąd kończy się karą dla operatora, powstaje silny bodziec do ukrywania problemów, przez co system bezpieczeństwa traci wiedzę o własnych słabościach.

Lotnictwo, medycyna i przemysł wysokiego ryzyka wypracowały dlatego koncepcję just culture, która pozwala odróżnić błąd od zaniedbania, naruszenia procedury czy świadomego działania ryzykownego. Podobne podejście będzie niezbędne w systemach AI: należy rozgraniczyć błąd operatora od błędu projektowego, niewłaściwej kalibracji modelu, wadliwej procedury czy organizacyjnie wymuszonego nadmiernego zaufania. Właśnie tutaj moral crumple zone okazuje się jednym z najważniejszych ostrzeżeń dla przyszłych organizacji. Człowiek nie może jednocześnie tracić kompetencji i czasu niezbędnego do kontroli, a wciąż zachowywać całkowitej odpowiedzialności za rezultat. Odpowiedzialność powinna podążać za kontrolą, wiedzą i realną możliwością zapobiegania zdarzeniom.

Jeśli producent wiedział o ograniczeniach modelu, lecz ich nie ujawnił, jego rola jest moralnie istotna. Jeśli organizacja zaprojektowała interfejs tak, by systematycznie promować automatyczną akceptację, również uczestniczy w odpowiedzialności. Z kolei jeśli operator świadomie zignorował czytelne ostrzeżenie, dysponując odpowiednimi kompetencjami i czasem na reakcję, ciężar przesuwa się w jego stronę. Etyka systemowa musi rekonstruować tę strukturę, zamiast szukać jednego podmiotu dla wygody instytucjonalnej. Odpowiedzialność rozproszona nie może jednak stać się odpowiedzialnością rozmytą. Stwierdzenie, że „wszyscy trochę odpowiadają", w praktyce często oznacza, że nikt nie odpowiada skutecznie.

Konieczne są zatem role przypisane ex ante: właściciel ryzyka, osoba odpowiedzialna za zatwierdzenie modelu, podmiot nadzorujący dane, operator wdrożenia, procedura audytu oraz mechanizm reakcji na incydent. W ten sposób odpowiedzialność nie jest dopisywana po katastrofie, lecz zostaje zaprojektowana przed uruchomieniem systemu. Taki sposób myślenia prowadzi do szerszej kategorii responsibility by design. W klasycznym safety engineering wiele błędów użytkownika interpretuje się nie jako nieuniknioną ludzką zawodność, lecz jako sygnał o źle zaprojektowanym systemie. Jeżeli interfejs systematycznie prowokuje do pomyłki, odpowiedź nie powinna ograniczać się do szkolenia pracownika, by „bardziej uważał". Podobnie w AI: jeśli użytkownicy konsekwentnie nadmiernie ufają rekomendacjom, należy zapytać, czy problem leży w psychologii człowieka, czy w sposobie prezentowania odpowiedzi. Człowiek posiada określone heurystyki i ograniczenia; etyczne projektowanie powinno uwzględniać je jako warunki brzegowe systemu, a nie traktować jako anomalie. Szczególnego znaczenia nabiera tutaj projektowanie perswazji – system generatywny może bowiem dostosowywać język, długość wyjaśnienia, ton emocjonalny i sposób argumentowania do konkretnego użytkownika.

Ochrona integralności poznawczej przed manipulacją systemową

Ta właściwość zwiększa użyteczność, lecz jednocześnie potęguje zdolność do wpływania na decyzje. Z etycznego punktu widzenia kluczowa staje się różnica pomiędzy pomocą epistemiczną a manipulacją poznawczą. Pomoc rozwija zdolność człowieka do trafnego rozumienia własnej sytuacji; manipulacja natomiast wykorzystuje mechanizmy psychologiczne, by skierować wybór w stronę interesu niejawnego lub niewspółmiernego do deklarowanego celu interakcji. Granica ta nie zawsze jest oczywista.

Lekarz przekonujący pacjenta do skutecznej terapii może stosować perswazję w jego interesie, jednak nadal musi szanować jego autonomię. System edukacyjny może motywować ucznia, lecz nie

powinien wykorzystywać podatności psychologicznej dziecka do realizacji komercyjnych celów dostawcy. Asystent zakupowy może faktycznie pomagać w wyborze, a może działać jak sprzedawca udający neutralnego doradcę. Im precyzyjniej AI modeluje jednostkę, tym ważniejsze staje się ujawnienie interesu, któremu służy dana rekomendacja. Problem alignment nie dotyczy więc wyłącznie zgodności abstrakcyjnych celów AI z wartościami ludzkości. Na poziomie codziennych systemów jest to również problem alignment of incentives: czy interes dostawcy, platformy, organizacji i użytkownika pozostaje spójny? Model może działać perfekcyjnie pod względem technicznym, a mimo to system jako całość może manipulować człowiekiem, jeśli celem biznesowym jest maksymalizacja czasu zaangażowania, liczby transakcji lub wymuszenie określonego zachowania. Etyka sprzężenia musi zatem obejmować nie tylko kod i model, ale ekonomię polityczną systemu. Transparentność powinna odnosić się nie tylko do tego, „jak działa AI", lecz przede wszystkim – dla kogo działa. Kto definiuje funkcję sukcesu? Kto czerpie korzyść z rekomendowanej decyzji? Kto ponosi koszt błędu?

Kto ma prawo modyfikować model? Kto kontroluje pamięć i historię użytkownika? W systemach hybrydowych pytania te stają się pytaniami o władzę poznawczą. Jeżeli AI uczestniczy w formowaniu zamiaru człowieka, ekonomiczny interes podmiotu kontrolującego technologię może zostać pośrednio włączony w proces kształtowania tego zamiaru. Właśnie dlatego etyka Hybrydowej Ludzkości powinna chronić integralność poznawczą. Pojęcie to jest szersze niż prywatność; podczas gdy ta odpowiada na pytanie, kto uzyskuje dostęp do danych o człowieku, integralność poznawcza pyta, kto i na jakich warunkach może ingerować w proces tworzenia przekonań, preferencji i decyzji. Różnica ta staje się kluczowa, gdy system nie tylko obserwuje zachowanie użytkownika, ale potrafi indywidualnie dostosować środowisko informacyjne tak, aby wywołać konkretną reakcję. Problem ten leży już na granicy neuroetyki, mimo że nie wymaga stosowania implantów.

Personalizowany system generatywny może znać styl reagowania użytkownika, jego rutyny, wcześniejsze wybory, podatność na określone argumenty oraz historię emocjonalną. Jeśli dodatkowo uzyska dostęp do danych biomedycznych lub sygnałów fizjologicznych, granica między rekomendacją a adaptacyjną modulacją stanu użytkownika ulegnie dalszemu zatarciu. Dlatego kolejny etap hybrydyzacji – BCI, neurotechnologie, biosensory i neuromodulacja – nie wprowadzi całkowicie nowego problemu etycznego, lecz zwiększy rozdzielczość i bezpośredniość kwestii, które już teraz występują w relacji człowieka z systemem informacyjnym. Zanim jednak przejdziemy do neurotechnologii, można sformułować ogólniejszą teorię legitymizowanego sprzężenia człowiek-AI. System jest tym bardziej etycznie dopuszczalny, im lepiej spełnia kilka powiązanych warunków: pozostaje responsywny wobec ludzkich racji; jego działanie daje się przypisać rozpoznawalnej strukturze odpowiedzialności; autonomia jest proporcjonalna do stawki i odwracalności;

niepewność jest komunikowana w sposób umożliwiający kalibrację; krytyczne decyzje mogą zostać zakwestionowane; istnieje możliwość zatrzymania i – tam, gdzie to możliwe – cofnięcia skutków; użytkownik zachowuje kompetencję i realną zdolność odmowy; a interes podmiotu kontrolującego system nie jest ukryty przed osobą, której zachowanie system współkształtuje. Nie są to niezależne zasady, lecz system wzajemnych zabezpieczeń.

Transparentność bez contestability może ujawnić niesprawiedliwość, ale jej nie naprawić. Contestability bez pamięci incydentów pozwoli odwołać pojedynczą decyzję, lecz system będzie powtarzał ten sam błąd. Human oversight bez czasu i kompetencji stanie się rytuałem. Rollback bez zapisu audytowego może cofnąć działanie, ale uniemożliwi ustalenie jego przyczyn. Ujawnianie niepewności bez pluralizmu źródeł nadal pozostawia użytkownika wewnątrz jednego modelu rzeczywistości. Bezpieczeństwo hybrydowej kognicji jest więc właściwością architektury wielu mechanizmów, a nie pojedynczego zabezpieczenia. Można z tego wyprowadzić jeszcze ważniejsze kryterium: dobra hybryda powinna umieć bezpiecznie odmawiać samej sobie pełnej autonomii. Inteligencja techniczna będzie coraz częściej oceniana przez zdolność do samodzielnego wykonywania zadań, lecz dojrzałość systemu może polegać również na prawidłowym rozpoznawaniu sytuacji, w których nie powinien działać.

W medycynie oznacza to eskalację nietypowego przypadku. W finansach – zatrzymanie transakcji o nietypowym profilu. W prawie – wskazanie konfliktu norm zamiast generowania pozornie jednoznacznej odpowiedzi. W agentowej AI – odmowę wykonania działania przy niejasnych uprawnieniach. Autonomia epistemicznie dojrzała zawiera mechanizm samoograniczenia. To prowadzi ponownie do AGE.

Suwerenność jako konstytucyjne zarządzanie zależnościami poznawczymi

System uczący się konsekwencji powinien nie tylko zwiększać skuteczność, ale także modyfikować własną politykę eskalacji. Jeśli wcześniejsze błędy wykazały, że określona klasa sytuacji jest problematyczna, późniejsza odpowiedź musi uwzględniać tę historię. Bezpieczeństwo staje się zatem częścią doświadczenia systemu. Dojrzałość agenta nie polega wyłącznie na ograniczaniu błędów dzięki wzrostowi kompetencji, lecz przede wszystkim na świadomości tego, gdzie te kompetencje się kończą.

Analogiczny obowiązek spoczywa na człowieku. Pokora epistemiczna nie oznacza rezygnacji z własnego osądu ani automatycznego podporządkowania się maszynie; jest raczej zdolnością do

rozpoznania ograniczeń zarówno ludzkiego, jak i sztucznego poznania. W tym ujęciu pokora epistemiczna służy jako mechanizm zachowania kontroli w systemie Fused Mind. Człowiek ma pozostać instancją nadającą kierunek nie dlatego, że zawsze wie więcej od AI, lecz ponieważ proces poznawczy wymaga normatywnego punktu odniesienia, który wykracza poza samą optymalizację. Nie oznacza to, że człowiek jest z definicji lepszym decydentem – historia polityki, administracji, medycyny czy wojny obfituje w przykłady ludzkich błędów i okrucieństwa. Meaningful human control nie może być fetyszem antropocentrycznym, wedle którego każdy udział człowieka automatycznie uszlachetnia działanie systemu. W strukturze hybrydy człowiek może stanowić jej słabszy komponent.

Sens kontroli nie tkwi w wierze w ludzką nieomylność, lecz w powiązaniu mocy decyzyjnej z odpowiedzialnością moralną, prawem i społeczną możliwością kwestionowania władzy. Maszyna nie posiada mandatu demokratycznego, statusu osoby odpowiedzialnej prawnie ani niezależnego uczestnictwa w normatywnej wspólnocie. Nawet gdy rekomendacja algorytmu statystycznie przewyższa ludzki osąd, decyzja o tym, jakie ryzyko zaakceptować, czyje prawa chronić i które wartości uznać za nadrzędne, pozostaje problemem politycznym i moralnym. Nie można zatem przejść od stwierdzenia, że model lepiej przewiduje rezultat, do wniosku, że model powinien decydować. Pomiędzy predykcją a decyzją znajduje się normatywna przestrzeń, której benchmark trafności nie rozstrzyga.

Przyszłość AI wymaga więc nie tylko alignmentu modeli, ale i alignmentu instytucji. Nawet idealnie dostrojony system zawiedzie, jeśli organizacja nagradza niewłaściwe cele, ukrywa błędy lub centralizuje władzę bez nadzoru. Bezpieczeństwo sztucznej inteligencji jest zatem problemem jednocześnie technicznym, psychologicznym, organizacyjnym i politycznym. Hybrydowa kognicja przekształca etykę AI w naukę o projektowaniu instytucji poznawczych. Kluczowe przesunięcie następuje w momencie przejścia od pytania „czy człowiek nadal naciska przycisk?” do pytania „czy struktura systemu pozwala społeczeństwu rozumieć, kwestionować i zmieniać sposób wytwarzania decyzji?”. To jest właściwa treść meaningful human control na poziomie cywilizacyjnym. Kontrola nie oznacza mikrosterowania każdą operacją superwydajnego systemu, lecz zachowanie zwierzchności nad regułami jego działania. Hybrydowa Ludzkość musi zachować autonomię, choć przyszły człowiek prawdopodobnie nie będzie osobiście wykonywał większości operacji poznawczych wpływających na jego życie.

Już dziś tego nie robi. Będzie jednak musiał zachować prawo i materialną możliwość ustalania granic delegacji, dostępu do własnej historii, zmiany infrastruktury oraz zatrzymywania działań przekraczających przyznane uprawnienia. Suwerenność w epoce AI nie oznacza zatem poznawczej

samowystarczalności, lecz zdolność konstytucyjnego zarządzania własnymi zależnościami poznawczymi.

Problem ten stanie się jeszcze bardziej radykalny, gdy przepływ między człowiekiem a maszyną wyjdzie poza ekran i mowę. Interfejsy mózg-komputer, neuroprotetyka czy neuromodulacja mogą przesunąć sprzężenie na poziom, w którym informacja techniczna wejdzie bezpośrednio do obiegu neuronalnego, a biologiczny stan użytkownika stanie się danym wejściowym systemu. Wówczas dotychczasowe pojęcia prywatności danych okażą się zbyt wąskie. Niezbędne staną się kategorie mental privacy, cognitive liberty, psychological continuity oraz mental integrity. Etyka sprzężenia przejdzie w etykę granicy samego podmiotu.

> Kiedy granica między biologicznym impulsem a algorytmiczną sugestią ostatecznie zniknie w interfejsach mózg-komputer, pytanie o kontrolę nad maszyną stanie się pytaniem o granice samego podmiotu. Czy w świecie totalnego sprzężenia będziemy jeszcze w stanie odróżnić własne pragnienia od precyzyjnie zaprogramowanej modulacji naszego stanu psychicznego? Suwerenność może okazać się ostatnim bastionem, którego nie da się zoptymalizować.

Inspiracje

[1] **"The AI Instinct"** Rana Gujral

Rana Gujral (ur. 1976) to amerykański przedsiębiorca, dyrektor naczelny i inwestor specjalizujący się w sztucznej inteligencji, sztucznej inteligencji kognitywnej (AI) i obliczeniach afektywnych. Najbardziej znany jest z roli dyrektora generalnego Behavioral Signals, firmy technologicznej dla przedsiębiorstw, która rozwija sztuczną inteligencję do rozpoznawania emocji i analizy mowy. Przed objęciem stanowiska dyrektora Behavioral Signals, Gujral założył firmę SaaS TiZE, która została przejęta przez Alchemy, i zajmował stanowiska kierownicze w takich firmach jak Logitech i Cricut, przyczyniając się do znaczących przekształceń korporacyjnych i rozwoju produktów. Praca Gujrala koncentruje się na styku ludzkiego podejmowania decyzji, uczenia maszynowego i kognitywistyki, gdzie rozwijał modele koncepcyjne, w tym sztuczne doświadczenie ogólne (AGE) i hybrydowe poznanie człowiek-maszyna. Jako częsty prelegent i komentator, porusza tematy etyki sztucznej inteligencji, inteligencji behawioralnej i transformacji w kierunku sztucznej inteligencji ogólnej.

Rana Gujral *(born 1976) is an American entrepreneur, business executive, and investor specializing in artificial intelligence, cognitive AI, and affective computing. He is best known for his leadership as the chief executive officer of Behavioral Signals, an enterprise technology company that develops emotion-recognition and speech-analytics artificial intelligence. Prior to leading Behavioral Signals, Gujral founded the enterprise SaaS company TiZE, which was acquired by Alchemy, and held senior executive positions at companies including Logitech and Cricut, contributing to significant corporate turnarounds and product development. Gujral's work focuses on the intersection of human decision-making, machine learning, and cognitive science, where he has advanced conceptual models including Artificial General Experience (AGE) and hybrid human-machine cognition. A frequent speaker and commentator, he addresses topics in AI ethics, behavioral intelligence, and the transition toward Artificial General Intelligence.*

Behavioral Signals

Literatura uzupełniająca

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] "Meaningful Human Control over Autonomous Systems: A Philosophical Account"

Filippo Santoni de Sio, Jeroen van den Hoven

2018 | Frontiers in Robotics and AI | DOI: 10.3389/frobt.2018.00015

Źródło

[2] "Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them"

Filippo Santoni de Sio, Giulio Mecacci

2021 | Philosophy & Technology | DOI: 10.1007/s13347-021-00450-x

Źródło

[3] "Meaningful human control as reason-responsiveness: the case of dual-mode vehicles"

Giulio Mecacci, Filippo Santoni de Sio

2019 | Ethics and Information Technology | DOI: 10.1007/s10676-019-09519-w

Źródło

[4] "Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight"

Ilse Verdiesen, Filippo Santoni de Sio, Virginia Dignum

2020 | Minds and Machines | DOI: 10.1007/s11023-020-09532-9

Źródło

[5] "Introduction to meaningful human control of artificially intelligent systems"

David Abbink, Daniele Amoroso, Luciano Cavalcante Siebert, Jeroen van den Hoven, Giulio Mecacci

2024 | Research Handbook on Meaningful Human Control of Artificial Intelligence Systems | DOI: 10.4337/9781802204131.00006

Źródło

[6] "Responsibility and innovation"

J. van den Hoven

2022 | Journal of Responsible Innovation | DOI: 10.1080/23299460.2022.2050570

Źródło

[7] **"Designing for Responsibility"**

Antonia Sattlegger, J. van den Hoven, Nitesh Bharosa

2022 | Digital Government Research | DOI: 10.1145/3543434.3543581

Źródło

Artykuły naukowe autora:

[8] **"The Nature of Human Cognition and AI's Approach to Mimicking It"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch02

Źródło

[9] **"AI in the Real World: Applications and Use Cases"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch03

Źródło

[10] **"The AI Instinct"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.fmatter

Źródło

[11] **"The Emergence of Artificial General Intelligence (AGI)"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch04

Źródło

[12] **"Introduction: Our Definition of “There” Is Wrong"**

RANA GUJRAL, KAI-FU LEE

2026 | DOI: 10.1002/9781394392155.ch01

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: 9a629d4e-09f7-45e2-aabe-7bf3570c5a55