This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/.

Kiedy systemy kontroli dziedziczą tę samą niepewność, którą mają regulować

Heather Grizzle · April 18, 2026

When AI guardrails are themselves probabilistic models, governance inherits the uncertainty it is meant to control. What real control requires.

Zarządzanie sztuczną inteligencją jest obecnie opisywane i sprzedawane poprzez to, co najłatwiej zwizualizować, a nie poprzez to, co najważniejsze do analizy.

W rozmowie dominują widoczne artefakty: panele kontroli polityk, warstwowe diagramy kontroli, oznaczone bariery ochronne oraz krok po kroku przedstawione przepływy wykonania, które sugerują, że żądanie jest rygorystycznie oceniane, zanim dojdzie do jakiegokolwiek działania. Te przedstawienia tworzą wrażenie kontroli, ponieważ przypominają tradycyjny nadzór nad systemem. Jednak przesłaniają one bardziej istotne pytanie o to, jaki rodzaj mechanizmu faktycznie dokonuje oceny wewnątrz tych warstw.

W większości obecnych architektur system AI generuje proponowane działanie, a propozycja ta jest kierowana przez sekwencję kontroli, która kończy się tym, że inny model decyduje, czy działanie powinno zostać wykonane. Ta warstwa oceny może być opisywana na różne sposoby w zależności od dostawcy lub ramy koncepcyjnej. Może być przedstawiana jako silnik polityk, mechanizm oceny ryzyka lub system rankingu pewności wzbogacony o reguły. Mimo tych różnic terminologicznych, podstawowa struktura pozostaje spójna. Model probabilistyczny ma za zadanie oceniać stosowność wyniku wygenerowanego przez inny model probabilistyczny.

Ten wybór projektowy ma istotne implikacje, gdy rozpatruje się go z perspektywy ryzyka i niezawodności. Systemy probabilistyczne z definicji nie produkują deterministycznych, powtarzalnych wyników w identycznych warunkach. Ich wyniki są kształto-

wane przez wzorce statystyczne wywodzące się z danych treningowych, a ich oceny są podobnie zdeterminowane przez wnioskowanie probabilistyczne, a nie stałe kryteria. Gdy taki system jest wykorzystywany do nadzorowania innego systemu tego samego typu, funkcja nadzorująca nie wprowadza fundamentalnie innej formy niezawodności. Zamiast tego powiela ten sam profil niepewności na wyższym poziomie abstrakcji.

Z punktu widzenia zarządzania tworzy to strukturę cykliczną. System, który ma egzekwować kontrolę, podlega tej samej zmienności, niestabilności przypadków brzegowych i wrażliwości na kontekst, co system, który nadzoruje. W rezultacie obecność dodatkowych warstw, kontroli czy nakładek niekoniecznie przekłada się na znaczącą redukcję ryzyka. Może zwiększać złożoność i może stwarzać pozory staranności, ale nie zmienia podstawowego zachowania systemu w sposób, który tworzyłby odrębną granicę kontroli.

> Wprowadza rekurencję, a nie kontrolę.

To rozróżnienie staje się jaśniejsze, gdy porówna się je z tym, jak zarządzanie funkcjonuje w innych dziedzinach wysokiego ryzyka. W systemach finansowych, na przykład, procesy zatwierdzania transakcji opierają się na z góry określonych regułach, dziennikach audytowych i kryteriach opracowanych przez ludzi, które można prześledzić i odtworzyć. W kontekstach inżynierii o krytycznym znaczeniu dla bezpieczeństwa systemy kontrolne są projektowane tak, aby działać z innymi właściwościami niezawodności niż systemy, które monitorują. Mechanizm nadzorujący jest celowo oddzielony zarówno pod względem funkcji, jak i zachowania, tak aby mógł stanowić stabilny punkt odniesienia w przypadku niepewności lub awarii.

Obecne wzorce zarządzania AI często nie ustanawiają tego rozdzielenia. Zamiast tego nakładają probabilistyczny osąd na probabilistyczną generację i traktują powstałą strukturę jako system kontroli. Takie podejście zakłada, że dodanie większej liczby kroków oceny w ramach tej samej klasy niezawodności przyniesie efekty w postaci zarządzania. W praktyce wprowadza ono rekurencję, a nie kontrolę, ponieważ każda warstwa zależy od tego samego rodzaju procesu rozumowania i jest podatna na te same rodzaje awarii.

Konsekwencja nie jest jedynie teoretyczna. Gdy system wytwarza wynik, który powoduje szkodę, pytanie, które następuje, nie dotyczy tego, czy w decyzję zaangażowanych było wiele warstw. Pytanie dotyczy tego, czy którakolwiek z tych warstw funkcjonowa-

ła jako prawdziwa kontrola, czyli działała z innym poziomem determinizmu, identyfikowalności i odpowiedzialności niż system, który nadzorowała. Jeśli każda warstwa dzieli tę samą probabilistyczną podstawę, to systemowi brakuje wyraźnego punktu, w którym można zakotwiczyć odpowiedzialność lub w sposób wiarygodny odtworzyć decyzje.

> **ROZSTRZYGAJĄCY TEST**
>
> Zarządzanie definiuje się poprzez wprowadzenie mechanizmu, który działa z innym profilem niezawodności niż system podlegający zarządzaniu.

Stanowisko wynikające z tej analizy jest jednoznaczne. Zarządzania nie definiuje liczba kontroli obecnych w systemie ani wizualna złożoność jego architektury. Zarządzanie definiuje się poprzez wprowadzenie mechanizmu, który działa z innym profilem niezawodności niż system podlegający zarządzaniu. Bez tego rozróżnienia dodatkowe warstwy nie stanowią kontroli. Rozszerzają tę samą niepewność na więcej komponentów.

Gdy warstwa kontrolna zachowuje się w taki sam sposób, jak system, który ma regulować, architektura nie osiąga zarządzania w sensownym znaczeniu. Wytwarza strukturę rekurencyjną, w której niepewność ocenia niepewność, a granica między działaniem a nadzorem zanika. To właśnie tam kryje się rzeczywiste ryzyko, niezależnie od tego, jak uporządkowany system wydaje się na pierwszy rzut oka.

> **TIPS**
>
> Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Jak cytować

Grizzle, H. M. (2026, April 18). Kiedy systemy kontroli dziedziczą tę samą niepewność, którą mają regulować. Novara Consulting Group. https://www.novaracg.com/pl/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/