Pomiar i zarządzanie wpływem społecznym: od metodologii danych do organizacyjnego uczenia się w modelu Matta Stevensona-Dodda

AUTOR

Fundacja Dobre Państwo

STRESZCZENIE / ABSTRACT

Artykuł analizuje wyzwania związane z pomiarem konstruktów latentnych, takich jak dobrostan, godność czy sprawczość, które nie posiadają naturalnych jednostek miary. Autor podkreśla konieczność stosowania zwalidowanych narzędzi psychometrycznych (np. skali WEMWBS) zamiast prostych ankiet autorskich, aby zapewnić rzetelność i porównywalność danych. Tekst zwraca uwagę na napięcie między standaryzacją a użytecznością praktyczną oraz wskazuje, że wybór instrumentu pomiarowego jest ściśle powiązany z przyjętą teorią człowieka. Kluczowym aspektem jest tzw. responsiveness – zdolność narzędzia do wykrywania rzeczywistych zmian wywołanych interwencją społeczną, co pozwala organizacjom przejść od intuicyjnego postrzegania efektów do dowodowego zarządzania wpływem.

The article analyzes the challenges associated with measuring latent constructs, such as well-being, dignity, or agency, which lack natural units of measurement. The author emphasizes the necessity of using validated psychometric tools (e.g., the WEMWBS scale) instead of simple proprietary surveys to ensure data reliability and comparability. The text highlights the tension between standardization and practical utility and indicates that the choice of a measurement instrument is closely linked to the adopted theory of the human person. A key aspect is so-called responsiveness—the tool's ability to detect actual changes caused by social intervention, allowing organizations to move from an intuitive perception of effects to evidence-based impact management.

Artykuł stanowi krytyczną analizę pomiaru wpływu społecznego, argumentując, że prawdziwa skuteczność organizacji nie wynika z akumulacji danych, lecz ze zdolności do ich interpretacji i wdrażania korekt strategicznych. Autor stawia tezę, że mierzenie konstruktów latentnych – takich jak godność czy dobrostan – wymaga odejścia od technokratycznego fetyszu liczb na rzecz pluralizmu metodologicznego, gdzie rygor psychometrii współistnieje z głębią analizy jakościowej i wsparciem AI. Kluczowym elementem tej architektury jest przejście od retrospektywnej ewaluacji do systemu Live Data, który nie służy jedynie obrazowaniu rzeczywistości na dashboardach, lecz tworzy organizacyjną pętlę uczenia się (double-loop learning). W tym ujęciu technologia ma za zadanie redukować obciążenie beneficjentów i przyspieszać przepływ informacji, jednak to kultura organizacyjna – zdolność do przyznania się do błędu i rewizji własnych założeń w obliczu dowodów – decyduje o tym, czy organizacja jest jedynie sprawnym administratorem danych, czy podmiotem realnie generującym zmianę społeczną.

SŁOWA KLUCZOWE

pomiar wpływu społecznego | model Matta Stevensona-Dodda | konstrukty latentne | dobrostan psychiczny | skala WEMWBS | psychometria w sektorze społecznym | response shift | attrition bias | Live Data | Theory of Change | evidence-responsive organisation | double-loop learning | triangulacja danych | heterogeneous treatment effects | organizacyjna epistemologia pokory

Pomiar konstruktów latentnych wymaga rygoru teoretycznego

Mierzenie tego, co nie daje się łatwo policzyć: dobrostan, doświadczenie, dane jakościowe i sztuczna inteligencja

Najtrudniejsza próba dla doktryny „mierz mniej, lecz lepiej” następuje wtedy, gdy przedmiotem zmiany przestaje być zdarzenie łatwe do jednoznacznego ustalenia – jak uzyskanie zatrudnienia, ukończenie szkolenia czy podpisanie umowy – a staje się nim **dobrostan, sprawczość,**

godność, bezpieczeństwo, zaufanie, poczucie przynależności, nadzieja albo zdolność do prowadzenia życia uznawanego przez człowieka za wartościowe.

Zjawiska te nie są mniej realne tylko dlatego, że nie posiadają naturalnej jednostki miary. Są jednak ontologicznie i metodologicznie odmienne od zmiennych fizycznych. Nie możemy przyłożyć "termometru godności" w takim sensie, w jakim mierzymy temperaturę, ponieważ przedmiot pomiaru jest konstruktem latentnym: o jego istnieniu i natężeniu wnioskujemy na podstawie odpowiedzi, zachowań, doświadczeń, relacji i kontekstu. Właśnie dlatego pomiar wpływu społecznego wkracza tutaj na teren psychometrii, fenomenologii, metodologii badań jakościowych, socjologii wiedzy i filozofii pomiaru.

Autor książki trafnie identyfikuje ten problem, wskazując, że w obszarach takich jak zdrowie psychiczne czy dobrostan prostą ankietę autorską warto niekiedy zastąpić narzędziem zwalidowanym, jako przykład podając Warwick-Edinburgh Mental Well-being Scale (WEMWBS). Jednocześnie zauważa on napięcie pomiędzy standaryzacją a użytecznością praktyczną: instrument naukowo solidny może być bardziej wymagający dla respondentów, a język akademicki lub kliniczny nie zawsze jest dostępny dla wszystkich grup. W tym napięciu zawiera się właściwie cała epistemologia „miękkich" rezultatów: **potrzebujemy porównywalności bez amputowania znaczenia, standaryzacji bez depersonalizacji oraz danych bez zamiany człowieka w zestaw pól formularza**.

W przypadku temperatury istnieje fizyczna wielkość o dobrze określonej strukturze pomiarowej. Z dobrostanem, poczuciem autonomii czy zaufaniem społecznym sytuacja jest bardziej złożona. Nie obserwujemy bezpośrednio „7,4 jednostki dobrostanu". Obserwujemy odpowiedzi osoby na serię pytań, jej funkcjonowanie, zachowanie i relacje, a niekiedy wskaźniki biologiczne lub administracyjne – i dopiero na tej podstawie wnioskujemy o stanie konstruktu teoretycznego. Z tego względu liczba powstała po zsumowaniu odpowiedzi nie jest samym dobrostanem, lecz **wynikiem instrumentu skonstruowanego według określonego modelu teoretycznego**.

Jeżeli zmienimy teorię dobrostanu, możemy zmienić pytania, strukturę skali i interpretację wyniku. Hedonistyczne ujęcie dobrostanu położy większy nacisk na pozytywny afekt i zadowolenie z życia; tradycja eudajmonistyczna – na rozwój, sens, autonomię i realizację możliwości; teoria capabilities – na realną swobodę osiągania wartościowych sposobów życia. To, co mierzymy, zależy więc częściowo od tego, **co wcześniej uznaliśmy za istotę badanego zjawiska**. Właśnie dlatego psychometria nie jest techniką neutralną wobec teorii człowieka.

Budowa skali rozpoczyna się od konceptualizacji, po której następuje operacjonalizacja, dobór pozycji oraz badanie struktury czynnikowej, rzetelności, trafności konwergencyjnej i

dyskryminacyjnej, czułości na zmianę oraz niezmienniczości pomiaru między grupami. Jeśli instrument działa statystycznie poprawnie w jednej populacji, nie oznacza to automatycznie, że posiada identyczne znaczenie w innej kulturze, grupie wiekowej czy sytuacji społecznej.

Dobrym przykładem jest wspomniana skala WEMWBS. Autor wskazuje, że pełna wersja obejmuje 14 pozycji, natomiast skrócona (SWEMWBS) siedem; obie posiadają udokumentowane właściwości psychometryczne i były stosowane w różnych kontekstach. Pełna wersja daje szerszy obraz dobrostanu, podczas gdy krótsza redukuje obciążenie respondenta przy zachowaniu dobrych właściwości skalowania. Obie są również wrażliwe na zmianę, co ma zasadnicze znaczenie w ewaluacji interwencji.

To ostatnie pojęcie – **responsiveness** – jest szczególnie ważne. Instrument może być rzetelny, czyli dawać stabilne i spójne wyniki, a jednocześnie słabo wykrywać zmianę istotną z perspektywy interwencji. W pomiarze wpływu nie wystarczy więc wiedzieć, czy skala „dobrze mierzy ludzi". Trzeba wiedzieć, czy potrafi wychwycić rzeczywistą zmianę, której program spodziewa się dokonać, oraz czy wielkość tej zmiany ma interpretowalne znaczenie.

Organizacje stosujące zwalidowane narzędzia mogą jednak ulec kolejnej formie fałszywego bezpieczeństwa. Walidacja skali rozwiązuje tylko część problemu: **rzetelny pomiar konstruktu nie dowodzi, że organizacja spowodowała zmianę tego konstruktu**. Jeśli wynik WEMWBS wzrósł po programie, mamy mocniejszy argument za tym, że mierzona zmienna psychologiczna rzeczywiście uległa zmianie, niż w przypadku prostego pytania: „czy czujesz się lepiej?".

Pułapki średniej i konieczność analizy dystrybucji efektów

Nadal jednak aktualne pozostają problemy atrybucji, deadweight, kontekstu i kontrfaktyczności, o których mowa było wcześniej. Można uznać, że ewaluacja wpływu składa się z trzech niezależnych warstw epistemicznych. Najpierw należy upewnić się, że mierzymy właściwą rzecz. Następnie – że mierzymy ją dostatecznie dobrze. Dopiero na końcu trzeba dowieść, że zmiana pomiaru jest związana przyczynowo z interwencją. Pominięcie dwóch pierwszych etapów prowadzi do pseudoprecyzji; zaniedbanie trzeciego – do nadmiernego przypisania sukcesu.

Równie istotne jest rozróżnienie pomiędzy istotnością statystyczną (statistical significance), wielkością efektu (effect size) a znaczącą zmianą (meaningful change). Bardzo duża próba może wykazać statystycznie istotną zmianę, która w praktyce ma znikomość znaczenie. Z kolei mała

organizacja może odnotować zmianę dużą i społecznie ważną, która jednak nie osiągnie klasycznego progu istotności z powodu niewielkiej liczebności grupy. Warwick publikuje informacje o minimalnie wykrywalnej zmianie dla skal WEMWBS i SWEMWBS, co pozwala interpretować różnice w sposób bardziej rzeczowy niż poprzez samo p-value.

Dla praktyki organizacyjnej płynie z tego kluczowa zasada: wynik skali nie powinien być samotną cyfrą. Należy go interpretować w powiązaniu z rozkładem wyników, wielkością efektu, zmianą indywidualną i kontekstem, a także z uwzględnieniem odsetka osób, u których nastąpiło pogorszenie oraz danych jakościowych wyjaśniających mechanizm zmiany. Średnia może poprawić się znacząco, podczas gdy podgrupa najbardziej potrzebująca nie korzysta z programu lub wręcz doświadcza regresu. Wpływ społeczny nie jest zatem jedynie kwestią średniej efektywności, lecz przede wszystkim dystrybucji efektów. W analizach społecznych średnia posiada ogromną siłę porządkującą, lecz bywa antropologicznie zdradliwa. „Przeciętny uczestnik" pozostaje bowiem jedynie konstruktem statystycznym.

Realni ludzie różnią się stanem początkowym, zasobami, traumami, dochodem, relacjami, zdrowiem, kompetencjami, pochodzeniem i zdolnością do korzystania z interwencji. Program może generować pozytywny średni wynik (outcome), przy jednoczesnym silnie zróżnicowanym rozkładzie rezultatów. Stąd coraz większe znaczenie zyskują tzw. heterogeneous treatment effects. Nie wystarczy ustalić, czy interwencja działa przeciętnie; trzeba pytać, dla kogo działa szczególnie dobrze, komu pomaga w niewielkim stopniu, u kogo nie działa wcale oraz czy istnieją grupy, którym szkodzi. W programach społecznych ta heterogeniczność nie jest zakłóceniem statystycznym, które należałoby „wygładzić" – często stanowi ona właściwy przedmiot wiedzy zarządczej.

Jeżeli program zwiększa dobrostan osób w relatywnie dobrej sytuacji początkowej, lecz nie dociera do tych w najtrudniejszych warunkach, średnia może wyglądać znakomicie, podczas gdy misja organizacji pozostaje niespełniona. Jeśli natomiast program skupia się na populacji wysokiego ryzyka, średnia zmiana może być niewielka, choć interwencja zapobiega dramatycznemu pogorszeniu stanu rzeczy. W pierwszym przypadku łatwy sukces może być iluzją sprawiedliwości; w drugim – pozorna przeciętność może ukrywać znaczący wpływ. Obnaża to również ograniczenia prostych rankingów organizacji: porównywanie surowych wyników bez uwzględnienia różnic w populacji wejściowej premiuje podmioty pracujące z łatwiejszymi przypadkami. W ekonomii edukacji, ochronie zdrowia i polityce społecznej prowadzi to do problemu risk adjustment – wynik należy interpretować w relacji do profilu potrzeb, nie przekształcając jednak korekty ryzyka w usprawiedliwienie niskich standardów.

Nawet najlepsza skala psychometryczna redukuje bogactwo doświadczenia do struktury odpowiedzi. Taka redukcja jest użyteczna, lecz nie wolno jej mylić z pełnią przeżyć. Człowiek może uzyskać ten sam wynik z bardzo różnych powodów. Dwie osoby o identycznym poziomie dobrostanu mogą różnić się źródłami cierpienia, posiadanymi zasobami, mechanizmem zmiany i tym, czego w danej chwili potrzebują.

Dane jakościowe jako rygorystyczny system dowodowy

W tym miejscu dane jakościowe zyskują znaczenie nieredukowalne do roli "kolorowego dodatku do statystyki". Zaktualizowany w 2026 roku brytyjski **Magenta Book** wprost wskazuje, że wywiady i grupy fokusowe pozwalają badać doświadczenia, intencje, postrzegany wpływ oraz mechanizmy zmiany. Odpowiadają one na pytania "jak?" i "dlaczego?", których sam monitoring ilościowy nie jest w stanie wyjaśnić. Z kolei jakościowe studia przypadków, obserwacja i etnografia pomagają osadzić rezultat w realnym kontekście działania. Co więcej, oficjalne brytyjskie ramy ewaluacji podkreślają przydatność metod jakościowych do identyfikowania zamierzonych i niezamierzonych efektów (outcomes), badania potrzeb różnych grup oraz wyjaśniania warunków, w których program odnosi sukces lub ponosi porażkę.

Wniosek ten ma kluczowe znaczenie dla modelu Stevensona-Dodda: dane jakościowe nie są przeciwieństwem "twardych dowodów". Mogą być twarde metodologicznie, nawet jeśli ich przedmiot nie przyjmuje formy liczby. Rygor badań jakościowych opiera się bowiem na innych mechanizmach niż reprezentatywność probabilistyczna. Fundamentem są tu: przejrzysty dobór przypadków, saturacja teoretyczna, konsekwencja kodowania, refleksyjność badacza, triangulacja oraz analiza przypadków przeczących. Istotne jest również posiadanie transparentnego codebooka, możliwość prześledzenia drogi od materiału źródłowego do interpretacji oraz rzetelne uzasadnienie generalizacji analitycznej. Brytyjskie ramy TIGER idą w tym kierunku jeszcze dalej, dążąc do zwiększenia reprodukowalności poprzez ujawnianie narzędzi, przewodników wywiadów i schematów kodowania również w analizach jakościowych.

Stevenson i Dodd zmierzają w tym kierunku, przeciwstawiając zwykłą anegdotę połączeniu **Data + Story**. W swojej analizie autorzy traktują dane ilościowe jako szkielet dowodowy, a narrację jakościową jako nośnik mechanizmu i kontekstu. Relację tę warto jednak pogłębić: nie chodzi o to, że liczba jest "prawdziwa", a historia jedynie "wyjaśniająca". Oba rodzaje danych są konstruowane i oba mogą być obarczone stronniczością.

Liczba zależy od definicji zmiennej, próby, brakujących danych, sposobu imputacji, modelu statystycznego i momentu pomiaru. Historia natomiast zależy od sformułowania pytania, pamięci

respondenta, relacji z badaczem, języka, sposobu transkrypcji oraz interpretacji. Różnica nie przebiega więc między obiektywnym a subiektywnym, lecz między **odmiennymi sposobami redukowania rzeczywistości**. Triangulacja jest wartościowa właśnie dlatego, że błędy obu perspektyw rzadko są identyczne.

Jeżeli skala dobrostanu rośnie, wywiady wskazują na wzrost poczucia sprawczości, uczestnicy podają spójne przykłady zmian w zachowaniu, a dane administracyjne potwierdzają stabilizację mieszkaniową lub zawodową – konwergencja tych niezależnych źródeł znacząco zwiększa wiarygodność interpretacji. Jeśli jednak źródła się rozchodzą, rozbieżność nie powinna być "naprawiana" poprzez wybór wygodniejszego wyniku; może ona sama w sobie stanowić najważniejsze odkrycie. Przykładowo: uczestnicy mogą deklarować wysoki poziom zadowolenia z programu, nie wykazując przy tym trwałej zmiany. Może to oznaczać, że program jest przyjemny, lecz nieskuteczny. Z drugiej strony mogą osiągać mierzalne rezultaty, a jednocześnie relacjonować doświadczenie utraty autonomii lub stygmatyzacji. Wówczas interwencja może być efektywna w jednym wymiarze i szkodliwa w innym. Tego rodzaju sprzeczności są szczególnie istotne, gdy przedmiotem działań jest człowiek, a nie maszyna.

Jednym z najtrudniejszych konstruktów jest godność. Można ją rozumieć normatywnie jako niezbywalny status osoby, którego nie da się "mieć więcej lub mniej", ale można ją również badać empirycznie jako **doświadczenie bycia traktowanym z szacunkiem**, poczucie autonomii, brak upokorzenia czy kontrolę nad codziennym życiem. Pomieszanie tych dwóch poziomów prowadzi do błędów konceptualnych.

Prawa człowieka nie zależą od wyniku ankiety, choć sposób ich doświadczania może być przedmiotem badania. Podobny problem dotyczy sprawiedliwości: możemy zmierzyć subiektywne poczucie sprawiedliwego traktowania, ale nie jest ono tożsame z normatywną sprawiedliwością instytucji. Ludzie mogą akceptować struktury niesprawiedliwe lub oceniać jako niesprawiedliwe systemy, które stosują uzasadnione zasady równego traktowania. Dane o percepcji są więc informacją o doświadczeniu społecznym, a nie ostatecznym sądem moralnym. Stevenson i Dodd przywołują w kontekście rzecznictwa trzy pytania związane z bezpieczeństwem, sprawiedliwym traktowaniem i możliwością rozwoju. Stanowią one interesujący **kompas aksjologiczny**, lecz nie powinny być mylone z kompletnym instrumentem walidowanym. Pytanie "czy czujesz się bezpieczniej?" może znakomicie otworzyć analizę doświadczenia, ale odpowiedź na nie zależy od punktu odniesienia, pamięci, aktualnego nastroju, zdarzeń społecznych i indywidualnej interpretacji pojęcia "bezpieczeństwa". Prostota pytania jest tu zaletą komunikacyjną, a nie gwarancją trafności psychometrycznej.

Pułapki pomiaru: od przesunięcia odpowiedzi po etykę danych

Response shift: kiedy sukces programu zmienia skalę, na której człowiek ocenia siebie

W pomiarze dobrostanu pojawia się subtelny problem znany jako **response shift**. Interwencja może bowiem zmieniać nie tylko stan człowieka, lecz także jego standardy, wartości i sposób interpretowania pytań. Osoba po terapii może rozumieć „dobrostan" zupełnie inaczej niż przed jej rozpoczęciem.

Człowiek odzyskujący autonomię często staje się bardziej krytyczny wobec swojej sytuacji, ponieważ rosną jego aspiracje. Paradoksalnie realna poprawa może więc nie przynieść proporcjonalnego wzrostu samooceny, a momentami towarzyszyć jej nawet spadek wyniku. Zjawisko to uderza w naiwną koncepcję ankiety jako okna bezpośrednio pokazującego rzeczywistość; odpowiedź jest zawsze relacją między doświadczeniem a układem odniesienia. Dlatego longitudinal measurement wymaga ostrożności interpretacyjnej, zwłaszcza w programach transformujących tożsamość, oczekiwania i poczucie możliwości. Ma to kluczowe znaczenie m.in. w aktywizacji osób doświadczających długotrwałego wykluczenia. Na początku programu oczekiwania mogą być niskie, przez co osoba ocenia swoją sytuację relatywnie dobrze w stosunku do znanego jej standardu. Jednak po poznaniu nowych możliwości wzrasta świadomość ograniczeń. W efekcie wynik „satysfakcji" może przejściowo spaść, choć capabilities realnie wzrosły. Bez danych jakościowych organizacja może błędnie uznać to za pogorszenie.

Szczególnie ważnym problemem w programach społecznych są brakujące dane. W typowej analizie brak ankiety traktuje się jako techniczny problem statystyczny, lecz w praktyce społecznej **nieodpowiadanie samo może być informacją**. Osoby w największym kryzysie częściej wypadają z programów: zmieniają adres, tracą kontakt, nie odbierają telefonów lub nie mają zasobów poznawczych, by wypełniać formularze. Jeżeli analiza obejmuje tylko tych, którzy pozostali w programie do końca, wynik może być systematycznie zawyżony. To klasyczny problem **attrition bias**. Badanie może wykazać wysoką poprawę wśród respondentów końcowych tylko dlatego, że osoby, którym program nie służył lub których sytuacja pogorszyła się najbardziej, zniknęły z próby. Dlatego stopa odpowiedzi, wzór braków i charakterystyka osób utraconych z obserwacji powinny być częścią samego wyniku, a nie przypisem technicznym.

W świecie impact management ma to również wymiar moralny. Najbardziej "nieme" statystycznie mogą być właśnie osoby, wobec których organizacja miała największą odpowiedzialność. System nastawiony wyłącznie na czyste dane może usuwać z obrazu tych, których życie jest najbardziej

chaotyczne. Autor książki zwraca uwagę na **survey burden** i możliwość ograniczania ankiet poprzez integrację istniejących danych. Zagadnienie to ma jednak nie tylko wymiar kosztowy. Sam proces zbierania informacji jest relacją społeczną i wpływa na respondenta. Wielokrotne pytanie człowieka o traumę, zadłużenie, przemoc czy zdrowie psychiczne nie jest neutralnym pobraniem bitów informacji.

Wrażliwy system powinien zatem stosować zasadę **data minimisation nie tylko jako wymóg ochrony prywatności, lecz jako zasadę etyczną ewaluacji**. Pytanie powinno zostać zadane tylko wtedy, gdy istnieje uzasadniona potrzeba. Jeżeli informacja znajduje się już w wiarygodnym źródle administracyjnym, ponowne pytanie może być zbędne; jeśli odpowiedź nie wpływa na decyzję ani badanie, obciążenie respondenta jest trudno uzasadnić. Z drugiej strony idea całkowitej „bezankietyzacji" wymaga sceptycyzmu. Dane administracyjne pokazują często to, co instytucja uznała za warte rejestrowania, a nie to, co człowiek uważa za ważne. Rejestr może odnotować fakt posiadania mieszkania, lecz nie powie, czy lokator czuje się w nim bezpieczny. Może potwierdzić zatrudnienie, ale nie wskaże, czy praca jest stabilna, godna i zgodna z możliwościami danej osoby. Eliminacja pytań bezpośrednich może więc obniżyć obciążenie, ale jednocześnie prowadzić do **epistemicznego wykluczenia doświadczenia**.

Dojrzały system nie powinien zatem pytać: „ankieta czy dane administracyjne?", lecz **które źródło najlepiej mierzy konkretny element Theory of Change przy najmniejszym uzasadnionym koszcie dla człowieka**. Autor książki formułuje efektowną tezę „Qualitative Strikes Back": modele językowe mają umożliwiać transkrypcję, klasyfikację tematów, analizę sentymentu i skalowanie analizy swobodnych wypowiedzi w stopniu wcześniej ekonomicznie niedostępnym. Kierunek ten jest widoczny, ale poziom pewności tej tezy wymaga skorygowania. Badania z 2025 r. dają obraz znacznie bardziej zniuansowany.

Studium poświęcone wykorzystaniu LLM do kodowania tekstu w analizie tematycznej wykazało, że modele traciły kompletność kodowania wraz ze wzrostem długości dokumentów i osiągały niską dokładność, co skłoniło badaczy do podkreślenia konieczności ludzkiego nadzoru i walidacji. Inne prace są bardziej optymistyczne: badania z sektora charity oraz prace nad skalowaniem kodowania jakościowego pokazują, że LLM mogą wspierać zarówno kodowanie dedukcyjne, jak i eksplorację większych korpusów – pod warunkiem zachowania jawnego workflow, kontroli badacza i świadomości interpretacyjnego charakteru procesu. Wniosek jest zasadniczy: AI nie „zamienia jakościowego w ilościowe". Raczej **obniża koszt niektórych operacji na materiale jakościowym**.

AI jako wsparcie, a nie zastępstwo interpretacji

Sztuczna inteligencja potrafi transkrybować, wyszukiwać podobne fragmenty, proponować kody, grupować wypowiedzi czy odnajdywać odstępstwa, pomagając w weryfikacji dużych zbiorów tekstu. Nie rozwiązuje ona jednak automatycznie problemu sensu, kontekstu, ironii, milczenia czy dynamiki władzy i relacji między wypowiedzią a sytuacją społeczną. Analiza jakościowa nie sprowadza się bowiem do klasyfikowania tekstu – jest przede wszystkim interpretacją. Kod „brak zaufania" może wynikać z dosłownej deklaracji, ale może też przejawiać się poprzez metaforę, unikanie odpowiedzi, niespójność, lokalny idiom lub historię znaną uczestnikom, lecz niewyrażoną wprost.

LLM znakomicie przetwarza język, lecz nie uczestniczy w społecznym polu wywiadu tak jak badacz. Stąd zasadność pojęcia skalowanej hermeneutyki. AI może rozszerzać zasięg analizy interpretacyjnej, ale nie powinna znosić jej hermeneutycznego charakteru. To badacz nadal definiuje problem i jednostkę analizy, tworzy lub zatwierdza codebook, interpretuje relacje między tematami oraz decyduje, co stanowi znaczący przypadek negatywny. Szczególne ryzyko wiąże się z automatyczną analizą sentymentu; przypisanie wypowiedzi etykiety „positive/negative/neutral" bywa bowiem bardzo słabym substytutem znaczenia społecznego. Zdanie „wreszcie przestałam oczekiwać pomocy" może zostać odczytane jako wyraz pozytywnej samodzielności lub negatywnej rezygnacji, zależnie od kontekstu.

W ewaluacji wpływu szczególnie niebezpieczne byłoby traktowanie automatycznego sentymentu jako miernika dobrostanu. Model klasyfikuje formę wypowiedzi, a nie dokonuje diagnozy stanu człowieka. Nawet jeśli statystycznie występuje tu związek, pozostaje on jedynie proxy podatnym na różnice kulturowe, językowe i indywidualne.

Najbardziej obiecujący model wykorzystania LLM w ewaluacji jakościowej nie polega na zastąpieniu badacza, lecz na stworzeniu relacji typu human-AI triangulation. Model może pełnić funkcję dodatkowego kodera, generować konkurencyjne interpretacje, wskazywać fragmenty niepasujące do dominującej kategorii lub pomagać w porównywaniu podgrup. Badacz natomiast kontroluje zasady, sprawdza próbki, analizuje rozbieżności i zachowuje odpowiedzialność za ostateczną interpretację. Taki układ jest poznawczo korzystniejszy niż prosta automatyzacja, ponieważ AI może działać jako adversarial collaborator.

Zamiast prosić model o potwierdzenie dominującego tematu, można nakazać mu wyszukiwanie wypowiedzi przeczących przyjętej interpretacji, alternatywnych mechanizmów lub grup, dla których narracja większości nie jest prawdziwa. Technologia wspiera wówczas falsyfikację zamiast produkcji eleganckich podsumowań. Takie zastosowanie odpowiada duchowi Theory of Change:

najcenniejsze dane to nie te, które potwierdzają program, lecz te, które pomagają rozróżnić konkurencyjne hipotezy. Wprowadzenie AI do analizy danych beneficjentów rodzi jednak problem epistemicznej asymetrii.

Ryzyka algorytmiczne i etyka danych w pomiarze wpływu

Modele językowe budowane są na ogromnych korpusach zastanej kultury. Kategorie, które wydają im się naturalne, często odzwierciedlają dominujące sposoby opisu człowieka. Jeśli organizacja dąży do odejścia od języka deficytowego, a model automatycznie interpretuje wypowiedzi według schematów patologizujących, technologia może nieświadomie przywracać dokładnie tę epistemologię, którą organizacja chciała porzucić. Podobne ryzyko dotyczy różnic językowych: wypowiedzi osób posługujących się dialektem, slangiem, językiem drugim, skrótem kulturowym lub stylem wynikającym z neuroatypowości mogą być kodowane inaczej niż język standardowy.

Jeżeli następnie kod AI zostanie przekształcony w wynik ilościowy, błąd językowy może stać się błędem alokacji zasobów. Dlatego system wykorzystujący AI powinien badać nie tylko ogólną zgodność kodowania z oceną człowieka, lecz również zgodność między grupami. Średnia dokładność (accuracy) może wyglądać dobrze, podczas gdy błędy koncentrują się w populacjach najbardziej marginalizowanych. To klasyczny problem fairness: model może być technicznie poprawny w ujęciu średnim, a jednocześnie społecznie niesprawiedliwy w rozkładzie błędów.

Dane jakościowe są wyjątkowo wrażliwe – im bogatsza narracja, tym łatwiej ujawnia ona informacje o zdrowiu, rodzinie, lokalizacji, relacjach, traumie czy statusie społecznym. Anonimizacja transkryptu nie zawsze sprowadza się do prostego usunięcia imienia i nazwiska; kombinacja zdarzeń życiowych może zidentyfikować osobę nawet bez danych bezpośrednich. Wykorzystanie AI dodatkowo komplikuje ten problem. Osoba, która zgodziła się na wywiad ewaluacyjny, mogła nie zakładać, że jej wypowiedź będzie przetwarzana automatycznie przez model, łączona z innymi źródłami lub używana do trenowania klasyfikatora. Dlatego innowacyjność technologiczna wymaga odpowiedniej legitymacji informacyjnej, a nie tylko zgodności technicznej.

Stevenson-Doddowska wizja "Human-Centric Technology" jest tutaj cenna: autorka pyta, czy AI i integracja danych mogą redukować obciążenie beneficjenta oraz czy zaoszczędzony czas wraca do relacyjnej pracy człowiek-człowiek. Należy jednak wprowadzić ostrzejsze kryterium: technologia skoncentrowana na człowieku nie jest nią tylko dlatego, że "oszczędza czas". Powinna także

zwiększać kontrolę jednostki nad danymi, ograniczać asymetrię informacyjną i nie zmieniać osoby w niewidzialny surowiec dla systemu analitycznego.

Jednym z najbardziej obiecujących kierunków jest łączenie danych zastanych. Stevenson-Dodd opisuje eksperyment związany z mapowaniem ubóstwa energetycznego poprzez połączenie danych o charakterystyce budynków, deprywacji obszarowej i zużyciu energii. Niezależnie od potrzeby szczegółowego zweryfikowania parametrów tego studium, kluczowa jest sama logika: organizacja nie zawsze musi pytać człowieka o zjawisko, które można trafniej rozpoznać w danych środowiskowych. To prowadzi do idei digital exhaust for social impact – wykorzystania śladów wytwarzanych przez systemy społeczne, które mogą informować o potrzebach lub rezultatach bez konieczności przeprowadzania dodatkowej ankiety. Dane o korzystaniu z energii, komunikacji, usług publicznych, edukacji czy zatrudnieniu mogą zmniejszyć koszty pomiaru i umożliwić obserwację longitudinalną. Jednocześnie korelacja w dużym zbiorze nie eliminuje problemu konstruktu: wysokie zużycie energii nie musi oznaczać ubóstwa energetycznego, gdyż może wynikać z wielkości gospodarstwa, ogrzewania elektrycznego, pracy zdalnej albo cech budynku.

Pluralizm metodologiczny jako ochrona przed technokratycznym paradoksem

Im bardziej decyzja dotyczy konkretnej osoby, tym większe znaczenie mają błędy typu false positive i false negative. Model, który świetnie nadaje się do mapowania ryzyka na poziomie obszaru, może być nieakceptowalny jako automatyczny klasyfikator uprawnień jednostki. To kluczowa różnica pomiędzy prediction for targeting a prediction for entitlement. Pierwsza z tych metod pomaga zdecydować, gdzie wysłać zespół badawczy lub uruchomić kampanię; druga zaś rozstrzyga o tym, kto otrzyma świadczenie.

Wzrost konsekwencji decyzji powinien przekładać się na wyższe wymagania dotyczące trafności, możliwości odwołania i ludzkiej kontroli. Najbardziej wymagającym polem w kontekście „trudnego do zmierzenia" wpływu jest rzecznictwo. Autor książki słusznie krytykuje traktowanie liczby spotkań z parlamentarzystami, raportów czy konferencji jako substytutu realnej zmiany prawnej i społecznej. W działaniach advocacy należy jednak unikać drugiej skrajności: oczekiwania, że jedynym prawdziwym outcome jest finalna zmiana prawa. Proces polityczny składa się z wielu etapów pośrednich: zmiany agendy, języka debaty, rozpoznawalności problemu, stanowisk ekspertów, koalicji interesariuszy, interpretacji prawnej, praktyki administracyjnej czy normy społecznej. Organizacja może wnieść istotny contribution, mimo że ustawa nie została jeszcze uchwalona. Może z kolei doprowadzić do zmiany ustawy bez trwałej zmiany praktyki.

Dlatego mierzenie rzecznictwa wymaga zastosowania outcome harvesting, process tracing, network analysis, media analysis oraz analizy dokumentów instytucjonalnych. Zamiast pytać wyłącznie o liczbę wykonanych działań, można rekonstruować sekwencję zmian: kiedy temat pojawił się w agendzie, kiedy zmieniła się narracja decydentów, które argumenty zostały przejęte, kto dołączył do koalicji, czy projekt aktu zawiera rozwiązania wcześniej promowane przez organizację oraz jakie inne czynniki mogły wywołać zmianę. W systemie politycznym wpływ przypomina więc bardziej falę przechodzącą przez sieć niż prostą linię A→B. Organizacja może inicjować, wzmacniać albo synchronizować proces, którego nie kontroluje. Właśnie tutaj pojęcie contribution jest szczególnie adekwatne.

Autor wprowadza także koncepcję Owning the Big Number: organizacja przyjmuje szeroki wskaźnik populacyjny jako strategiczny punkt odniesienia, zamiast mierzyć jedynie własne działania. Podejście to jest intelektualnie atrakcyjne, ponieważ wymusza orientację na problem, a nie na samą organizację, jednak niesie ze sobą ogromne ryzyko attribution.

Jeżeli organizacja zajmująca się aktywnością fizyczną osób z niepełnosprawnościami przyjmuje krajową lukę aktywności jako „big number", postępuje sensownie, o ile traktuje go jako north star metric. Postęp w skali kraju nie może jednak być automatycznie raportowany jako jej własny impact. Makrowskaźnik jest bowiem produktem polityk publicznych, trendów społecznych, infrastruktury, technologii, gospodarki i działań wielu podmiotów.

„Owning" powinno zatem oznaczać przyjęcie odpowiedzialności strategicznej za orientację na problem, a nie epistemiczne przywłaszczenie rezultatu. Organizacja może pytać, czy jej contribution jest dostatecznie duży w stosunku do skali problemu, ale nie może z faktu zmiany makro wyprowadzać wyłącznej zasługi. Najważniejsza ewolucja całego obszaru polega być może na przejściu od measurement of beneficiaries do learning with people. Tradycyjny system traktuje beneficjenta jako źródło danych: odpowiada on na ankietę, generuje outcome i pojawia się w bazie. Bardziej dojrzały system widzi w nim współinterpretatora rezultatów. Dane mogą pokazać wzrost wyniku o cztery punkty; człowiek natomiast powie, co te cztery punkty znaczyły w jego życiu – czego skala nie uchwyci – i dlaczego program zadziałał. Może również zakwestionować definicję sukcesu przyjętą przez organizację. Tym samym lived experience wraca nie tylko na początku, przy projektowaniu purpose, lecz także na końcu – przy interpretowaniu wpływu. Zamyka to ważną pętlę epistemiczną. Organizacja nie tylko mierzy zmianę w człowieku; człowiek uczestniczy w ocenie, czy organizacja właściwie rozumie zmianę. Taka zasada chroni przed technokratycznym paradoksem, w którym system wie coraz więcej danych o osobie, a coraz mniej rozumie jej sytuację.

Z analizy wynika, że spór między podejściem ilościowym a jakościowym jest w dużej mierze źle postawiony. Metody nie powinny konkurować o status jedynego prawdziwego poznania, gdyż

odpowiadają na różne klasy pytań. Dane ilościowe są szczególnie przydatne do określania skali, rozkładu, różnic, trendów i efektów. Dane jakościowe z kolei silnie oddziałują w rekonstruowaniu znaczeń, mechanizmów, doświadczenia i kontekstu. Dane administracyjne mogą ukazywać zachowania bez dodatkowego obciążania ludzi, psychometria pozwala operacjonalizować konstrukty latentne, a etnografia odkrywać różnicę między usługą zaprojektowaną a rzeczywiście przeżywaną. AI może natomiast zwiększać przepustowość analityczną.

Żadna z tych metod sama w sobie nie stanowi kompletnego systemu wiedzy. Aktualny Magenta Book przyjmuje właśnie taki pluralizm: metody ilościowe, jakościowe, eksperymentalne, theory-based i studia przypadków są dobierane do konkretnego pytania ewaluacyjnego. Od 2026 roku przewodnik ten jeszcze silniej podkreśla bieżące uczenie się i poprawę programu, zamiast redukować ewaluację do rekonstrukcji tego, „co się wydarzyło". To zbliża nowoczesną ewaluację do intuicji Stevenson-Dodda, a jednocześnie ją koryguje. Nie potrzebujemy wyboru między „prostymi, ludzkimi miarami" a „akademickim aparatem". Potrzebujemy proporcjonalnego aparatu badawczego, którego stopień komplikacji odpowiada ryzyku decyzji, niepewności teorii zmiany i znaczeniu społecznemu programu.

Automatyzacja wspierająca sprawczość i przejście ku architekturze Live Data

W kontekście AI najważniejszą zasadą nie powinna być automatyzacja wszystkiego, co możliwe, lecz automatyzacja tego, czego proces ten zwiększa ludzką sprawczość i jakość decyzji. Jeśli AI redukuje czas ręcznego kodowania, pozwalając badaczowi poświęcić więcej uwagi interpretacji, realnie podnosi wartość pracy. Jeżeli jednak automatycznie zamienia wypowiedzi ludzi w sentymenty, eliminując potrzebę kontaktu, może zwiększać skalę działań kosztem głębi rozumienia.

Postulat Stevenson-Dodda, aby oszczędności technologiczne wracały do relacji human-to-human, jest ważniejszy niż sama technologia. Można go potraktować jako szczególną zasadę alokacji: automatyzuj procedurę, nie odpowiedzialność; redukuj administrację, nie relację; zwiększaj analizę, nie pozbawiaj człowieka prawa do interpretacji własnego doświadczenia. W tym miejscu przebiega granica między AI jako instrumentem wpływu a AI jako nową warstwą biurokracji. Pierwsza zwiększa zdolność organizacji do rozumienia świata; druga generuje więcej raportów, klasyfikacji i dashboardów, niż ludzie są w stanie wykorzystać. Technologia może zatem leczyć lub pogłębiać Stevenson-Doddowski problem „Do": sprawnie produkować analitykę, która niczego nie zmienia.

Paradoks ten można ująć następująco: im bardziej subtelny jest przedmiot, tym większego rygoru wymaga jego pomiar. Trudność nie usprawiedliwia dowolności. Stwierdzenie, że „dobrostan jest złożony", nie może oznaczać zadania jednego, przypadkowego pytania. Fakt, iż „godność jest niematerialna", nie sprawia, że każda narracja o niej jest równie trafna. Podobnie „złożoność wpływu społecznego" nie oznacza, że nie podlega on weryfikacji. Jednocześnie rygor nie musi oznaczać przymusowej kwantyfikacji.

Czasem najbardziej rygorystycznym rozwiązaniem okaże się dobrze zaprojektowany wywiad, analiza procesu, obserwacja etnograficzna lub studium przypadku. Twardość dowodu nie wynika z obecności cyfry, lecz ze spójności między pytaniem, metodą, materiałem i siłą wniosku. Dobra organizacja typu „Difference" nie mierzy człowieka dlatego, że może. Buduje system, w którym każda informacja pełni funkcję w Theory of Change, metoda odpowiada naturze badanego konstruktu, a dowody ilościowe i jakościowe wzajemnie się testują. Technologia ma tu zwiększać zdolność interpretowania, nie przywłaszczując doświadczenia. Jej najwyższą ambicją nie jest przełożenie życia społecznego na dashboard, lecz stworzenie obrazu rzeczywistości dostatecznie wiarygodnego, by organizacja mogła zmienić własne działanie, kiedy rzeczywistość przeczy założeniom.

Samo posiadanie dobrych danych nie tworzy jednak organizacji uczącej się. Dane mogą zostać zebrane i zwizualizowane, a mimo to nic nie zmienić. Przechodzę zatem do kwestii danych na żywo, sprzężeń zwrotnych i organizacyjnego uczenia się. Zestawię koncepcję Stevenson-Dodda z cybernetyką, teorią single-loop i double-loop learning Argyrisa i Schöna oraz adaptive management, wskazując różnicę między dashboardem, który jedynie obrazuje rzeczywistość, a systemem, który potrafi dzięki niej zmieniać własne reguły działania.

Przejście od retrospektywnej ewaluacji do danych dostępnych w krótkich cyklach nie jest wyłącznie zmianą techniczną – to zmiana epistemologii zarządzania. Organizacja przestaje traktować wiedzę o własnym wpływie jako produkt sporządzany po zakończeniu działań, a zaczyna ujmować ją jako element samego procesu. W modelu tradycyjnym interwencja zostaje zaprojektowana i wdrożona, a dopiero na końcu oceniona; w modelu adaptacyjnym obserwacja rezultatu jest wpisana w przebieg programu i może modyfikować jego kierunek.

Stevenson-Dodd przeciwstawia sobie „lusterko wstecznego" i architekturę Live Data: dane operacyjne trafiają przez systemy CRM i interfejsy API do warstwy analitycznej, co umożliwia szybszą korektę działań. Nie chodzi jednak o prostą przewagę szybkości nad raportem. Istotniejsza jest zmiana funkcji informacji – z dokumentowania przeszłości na tworzenie sprzężenia zwrotnego pomiędzy działaniem, obserwacją, interpretacją i kolejną decyzją.

W języku cybernetyki organizację można potraktować jako system regulacyjny. Dysponuje ona stanem pożądanym, obserwuje stan rzeczywisty, identyfikuje odchylenie i podejmuje działanie korygujące. Jeśli celem programu jest trwałe zatrudnienie uczestników, samo prowadzenie szkoleń nie wystarcza do regulacji systemu. Potrzebna jest informacja o tym, czy ludzie znajdują pracę, jak długo ją utrzymują i gdzie pojawiają się systematyczne bariery. Dopiero świadomość rozbieżności między oczekiwaniami a rzeczywistością pozwala organizacji zmieniać zachowanie. Pomiar pełni tu funkcję sensora: jego znaczenie nie polega na samym rejestrowaniu świata, lecz na tym, że obserwacja wchodzi w pętlę sterowania.

Taka perspektywa precyzyjniej wyjaśnia postulat Live Data. Stevenson-Dodd wskazuje, że zamiast czekać na roczny raport, zarząd może dzięki bieżącym danym obserwować pełną ścieżkę beneficjenta – od wejścia do programu, przez trudności i ewentualną rezygnację, aż po trwały rezultat – co pozwala szybciej modyfikować lub zamykać nieskuteczne działania.

Dane stają się sprzężeniem zwrotnym tylko w kulturze uczenia dwupętlowego

Wartość tej architektury nie wynika jednak z samej częstotliwości odświeżania ekranu. Dashboard aktualizowany co godzinę może być równie bezużyteczny poznawczo, co pięćsetstronicowy raport roczny, jeśli prezentuje źle dobrane wskaźniki lub gdy nikt nie ma uprawnień do zmiany programu. Dane stają się sprzężeniem zwrotnym dopiero wtedy, gdy istnieje mechanizm instytucjonalny zdolny na nie odpowiedzieć.

W tym miejscu model Stevenson-Dodda można owocnie zestawić z klasycznym rozróżnieniem Chrisa Argyrisa i Donalda Schöna pomiędzy **single-loop learning** a **double-loop learning**. Argyris definiował uczenie organizacyjne jako proces wykrywania i korygowania błędów. W uczeniu jednopoziomowym organizacja modyfikuje swoje działania, by skuteczniej osiągać istniejący cel, nie kwestionując przy tym podstawowych norm i założeń. Z kolei w uczeniu dwupętlowym samo pojawienie się błędu staje się impulsem do zakwestionowania reguł, polityk i celów, które doprowadziły do określonego sposobu działania. Rozróżnienie to ma kluczowe znaczenie dla zarządzania wpływem, ponieważ większość współczesnych dashboardów wspiera głównie uczenie pierwszej pętli. Jeśli organizacja zauważy spadek frekwencji na zajęciach, może zmienić godzinę spotkań, kanał komunikacji lub sposób przypominania uczestnikom. Jeśli zatrudnienie absolwentów okaże się niższe od oczekiwanego, może zmodyfikować moduł szkoleniowy.

Są to klasyczne korekty **single-loop**: cel pozostaje nienaruszony, zmienia się jedynie droga do jego osiągnięcia. Uczenie dwupętlowe pojawia się dopiero wtedy, gdy organizacja zaczyna pytać, czy samo zatrudnienie jest właściwą definicją sukcesu, czy wybrana populacja jest trafna, czy założony mechanizm interwencji ma sens, czy wymaganie aktywizacji nie jest błędnie dopasowane do potrzeb części uczestników lub czy program nie próbuje rozwiązywać problemu indywidualnego, który w rzeczywistości ma źródło strukturalne. Różnicę między tymi poziomami można sprowadzić do pytania: „czy robimy to dobrze?" kontra „czy robimy właściwą rzecz?". Pierwsze dotyczy wykonania i efektywności instrumentalnej; drugie odnosi się do teorii działania, wartości, priorytetów i samej konstrukcji celu.

Organizacja może perfekcyjnie realizować **single-loop learning** i jednocześnie przez lata optymalizować program zbudowany na błędnej hipotezie. Każda kolejna wersja procedury staje się sprawniejsza, lecz niekoniecznie bardziej trafna. Właśnie dlatego Live Data bez zdolności do **double-loop learning** może prowadzić do szybszej optymalizacji błędu. Argyris zwracał uwagę na mechanizmy obronne utrudniające uczenie dwupętlowe. Normy, relacje władzy i reputacyjne koszty przyznania się do pomyłki mogą sprawić, że pracownicy dostrzegą problem wcześniej niż kierownictwo, lecz informacja ta nie przełoży się na zmianę podstawowych reguł. Przykład przedsiębiorstwa kontynuującego produkcję wadliwego produktu mimo wiedzy menedżerów pokazuje, że brak uczenia organizacyjnego często wynika nie z braku danych, lecz z blokady komunikacyjnej i obrony istniejących założeń. W kontekście pomiaru wpływu jest to kluczowe: organizacja może posiadać znakomity dashboard i jednocześnie nie mieć kultury pozwalającej potraktować niekorzystny wynik jako informację, a nie zagrożenie. Dlatego do architektury Live Data należy dodać trzeci element – **deutero-learning**, czyli uczenie się tego, jak organizacja się uczy.

Nie chodzi tu już o korektę działania ani samo zakwestionowanie założenia, lecz o refleksję nad zdolnością wykrywania i naprawiania błędów. Organizacja zaczyna badać, które informacje trafiają do zarządu, jakie sygnały są tłumione, kto może zakwestionować dominującą interpretację, czy anomalie są wyjaśniane, czy raczej racjonalizowane oraz ile czasu mija między wykryciem problemu a zmianą decyzji. Literatura rozwijająca myśl Argyrisa i Schöna traktuje ten poziom jako uczenie o procesach uczenia, obejmujące kulturę, normy, struktury i sposoby organizowania refleksji.

W tym świetle dojrzały system **impact management** powinien monitorować nie tylko rezultaty programu, lecz również jakość własnej reakcji na te rezultaty. Jeśli organizacja przez pół roku obserwuje pogarszający się outcome, a mimo to niczego nie zmienia, problemem przestaje być sam wynik – pojawia się problem governance.

Jeżeli dane wskazują, że jedna grupa systematycznie wypada z programu, lecz system raportowania traktuje to wyłącznie jako „brakujące dane", mamy do czynienia z porażką epistemiczną. Gdy pracownicy pierwszej linii wiedzą, że kryterium kwalifikacyjne jest wadliwe, ale nie istnieje procedura przekazania tej informacji na poziom strategiczny, architektura danych staje się jedynie pozorem uczenia. Autor książki wprost wskazuje, że Fundamenty Wpływu i Theory of Change nie powinny być traktowane dogmatycznie, lecz konfrontowane z rzeczywistością poprzez regularną pętlę feedbacku. Liderzy mają sprawdzać, czy beneficjenci potwierdzają zakładane mechanizmy, czy zespoły rozumieją **purpose** oraz czy organizacja potrafi otwarcie komunikować nieskuteczność i porażki. Intuicja ta odpowiada dokładnie **double-loop learning**: Theory of Change przestaje być manifestem i staje się hipotezą operacyjną poddawaną rewizji. Warto jednak wyraźnie odróżnić częstotliwość danych od częstotliwości decyzji. W zarządzaniu technologicznym często zakłada się, że im bardziej „real time" są dane, tym lepiej. Nie zawsze jest to prawdą.

Świadome zarządzanie opóźnieniem zamiast fetyszu natychmiastowości

Niektóre rezultaty zmieniają się powoli, wykazują sezonowość, naturalną zmienność lub silny komponent losowy. Zbyt szybka reakcja na krótkookresowe wahania może prowadzić do **overcontrol** – nadmiernego sterowania systemem na podstawie szumu zamiast sygnału. Zarząd obserwujący codzienną fluktuację dobrostanu uczestników mógłby wprowadzać ciągłe korekty programu, mimo że trend kwartalny pozostaje stabilny. W statystycznej kontroli procesów znany jest problem **tampering**: ingerowanie w proces po każdej naturalnej fluktuacji może zwiększać jego wariancję zamiast ją redukować.

Dlatego architektura Live Data wymaga wprowadzenia pojęcia odpowiedniego rytmu poznawczego. Dane finansowe dotyczące płynności mogą wymagać aktualizacji codziennej, informacja o frekwencji – tygodniowej, a outcome zatrudnieniowy – miesięcznej lub kwartalnej. Zmiana strukturalna w jakości życia wymaga natomiast jeszcze rzadszej obserwacji. Nie istnieje jedna optymalna częstotliwość; powinna ona wynikać z dynamiki badanego procesu, kosztu reakcji, prawdopodobieństwa fałszywego alarmu oraz realnych możliwości modyfikacji interwencji. Można to ująć jako problem latency. System obarczony jest kilkoma rodzajami opóźnień: od wystąpienia zdarzenia do jego rejestracji, od rejestracji do analizy, od analizy do zauważenia przez decydenta, od zauważenia do decyzji oraz od decyzji do zmiany praktyki. Organizacja może technicznie dysponować „danymi w czasie rzeczywistym", a w rzeczywistości borykać się z kilkumiesięcznym opóźnieniem decyzyjnym. Z drugiej strony całkowite skrócenie tych etapów nie zawsze jest

wskazane, gdyż interpretacja niektórych zjawisk wymaga czasu i kontekstu. Dojrzałość polega zatem na świadomym zarządzaniu opóźnieniem, a nie na fetyszu natychmiastowości.

Problem ten wiąże się z pojęciami obserwowalności i sterowalności z teorii systemów. System jest poznawczo użyteczny wtedy, gdy na podstawie dostępnych sygnałów można wystarczająco precyzyjnie wnioskować o jego stanie. Jest natomiast sterowalny, gdy istnieją działania pozwalające doprowadzić go do pożądanego stanu.

Organizacja społeczna może cechować się wysoką obserwowalnością przy niskiej sterowalności: może bardzo dokładnie wiedzieć, że wzrasta bezdomność, lecz nie dysponować zasobami ani kompetencjami, by zmienić politykę mieszkaniową. Może również mierzyć się z odwrotnym problemem: posiadać narzędzia interwencji, ale brakować jej danych pozwalających określić, gdzie i kiedy je zastosować. Z tego powodu jednym z najpoważniejszych błędów systemów dashboardowych jest mieszanie sygnałów informacyjnych z dźwigniami decyzyjnymi. Nie każdy wskaźnik wart obserwacji znajduje się w sferze kontroli organizacji. Wzrost regionalnego bezrobocia może być kluczową informacją kontekstową, jednak mała fundacja nie jest w stanie nim sterować.

Live Data jako system diagnostyczny dla adaptacyjnego zarządzania

System powinien zatem odróżniać wskaźniki outcome, które organizacja współtworzy, od wskaźników kontekstu, pomagających interpretować wynik. Bez tego rozgraniczenia zarząd może być rozliczany z procesów pozostających poza jego realną sprawczością lub przeciwnie – przypisywać sobie sukcesy będące wynikiem korzystnej zmiany otoczenia. Stevenson-Dodd przedstawia Live Data Storytelling jako połączenie koncentracji na kilku kluczowych wskaźnikach, czytelnej architektury informacji, automatycznej synchronizacji danych oraz przejrzystego ujawniania nieskuteczności. Należy jednak dodać piąty wymiar: interpretowalność przyczynową. Dashboard nie może jedynie informować o spadku wskaźnika; musi umożliwiać zejście do poziomu konkretnego segmentu, czasu, lokalizacji i mechanizmu. W przeciwnym razie wykres staje się sygnalizatorem bez diagnostyki. Tu właśnie pojawia się znaczenie drill-down.

Jeśli ogólny outcome ulega pogorszeniu, zarząd powinien mieć możliwość sprawdzenia, czy problem dotyczy całej populacji, konkretnego regionu, grupy wiekowej, nowego sposobu rekrutacji, określonego zespołu czy może zmiany sezonowej. Nie chodzi o to, by członkowie rady nadzorczej stawali się analitykami danych, lecz by system pozwalał przejść od pytania „co się dzieje?” do pytań:

„gdzie to się dzieje?” oraz „jaka hipoteza może to wyjaśniać?”. Bez tej funkcji dashboard może wzmacniać reakcje impulsywne zamiast wspierać proces uczenia. Równie istotne jest rozróżnienie leading indicators od lagging indicators. Outcome końcowy bywa wskaźnikiem opóźnionym – stabilność zatrudnienia przez sześć miesięcy można ocenić dopiero po pół roku. Jeśli organizacja czeka wyłącznie na taki wynik, jej reakcja będzie z natury spóźniona.

Wymagane są zatem wcześniejsze sygnały powiązane z Theory of Change: frekwencja, zaangażowanie, nabywanie kompetencji, jakość relacji z mentorem czy liczba rozmów rekrutacyjnych. Zmienne te nie stanowią samego wpływu, ale mogą pełnić funkcję wczesnych ostrzeżeń, o ile istnieją dowody, że rzeczywiście poprzedzają outcome. To moment, w którym łatwo ponownie pomylić output z impact. Leading indicator jest użyteczny tylko wtedy, gdy traktuje się go jako sygnał prognostyczny, a nie substytut rezultatu. Jeśli organizacja uzna, że sama frekwencja staje się sukcesem tylko dlatego, iż koreluje z zatrudnieniem, wraca do prymitywnego modelu „Do”. Dobrze zaprojektowany system musi utrzymywać tę różnicę: jedne wskaźniki służą adaptacji, inne rozliczają rezultat, a jeszcze inne opisują kontekst.

Nowoczesna literatura o zarządzaniu publicznym wyraźnie przesuwa się w stronę takich pętli adaptacyjnych. OECD wskazuje, że monitoring i ewaluacja nie powinny jedynie rozliczać polityki po fakcie, lecz poprzez feedback loops łączyć obserwowane outcomes i impacts z kolejnymi decyzjami, wzmacniając funkcję uczenia. W publikacji Science, Technology and Innovation Outlook 2025 OECD mówi wręcz o potrzebie przejścia od ewaluacji ex post ku ewaluacji służącej „real-time learning” – szczególnie w dynamicznych środowiskach, gdzie wiedza docierająca po zakończeniu programu może być poznawczo poprawna, ale operacyjnie bezużyteczna.

Stevenson-Dodd nie stoi zatem poza głównym nurtem współczesnego zarządzania publicznego; jego praktyczna intuicja wpisuje się w szerszy zwrot ku adaptive governance. Należy jednak zachować istotne rozróżnienie między adaptive management a nieustannym improwizowaniem. Adaptacyjność nie polega na częstej zmianie kierunku, lecz na korekcie uzasadnionej dowodem. Organizacja powinna posiadać wcześniej określone progi, warunki przeglądu i zasady interpretacji sygnałów. Jeśli każda niekorzystna dana wywołuje zmianę programu, traci się stabilność niezbędną do oceny skuteczności interwencji. Adaptacyjne zarządzanie wymaga więc paradoksalnie zarówno elastyczności, jak i dyscypliny metodologicznej.

Logikę tę można porównać z iteracją naukową: hipoteza jest formułowana, testowana, a wyniki interpretowane, co może prowadzić do jej modyfikacji. Badacz nie zmienia jednak hipotezy po każdej pojedynczej obserwacji, jeśli nie ma podstaw uznać jej za diagnostyczną. Analogicznie organizacja potrzebuje reguł aktualizacji przekonań. Bayesowska interpretacja impact management

rozwinięta wcześniej znajduje tutaj praktyczne zastosowanie: sygnały napływają nieustannie, lecz różnią się mocą dowodową i nie każdy z nich powinien w równym stopniu zmieniać strategię.

Szczególną uwagę należy poświęcić problemom concept drift i data drift. W systemach działających przez lata relacja między miernikiem a zjawiskiem może ewoluować. Wzorzec skuteczności sprzed pięciu lat może przestać być trafny z powodu zmian na rynku pracy, zachowań użytkowników czy otoczenia prawnego. AI i modele predykcyjne zwiększają ryzyko reagowania na świat opisany danymi historycznymi, przy jednoczesnym przeoczeniu zmiany mechanizmu generującego te dane.

Dlatego dojrzały system Live Data powinien monitorować nie tylko wartości wskaźników, lecz także stabilność ich znaczenia. Jeśli ta sama wartość przestaje oznaczać to samo, dashboard może pozostawać technicznie poprawny i jednocześnie zarządczo błędny. Jest to szczególnie istotne w sektorze społecznym, gdzie populacje i konteksty zmieniają się szybciej niż definicje KPI.

Stevenson-Dodd przywołuje YMCA England and Wales jako przykład systemu federacyjnego, w którym rozproszone lokalne dane zintegrowano poprzez wspólną infrastrukturę chmurową i API. Umożliwiło to kompleksową analizę obszarów takich jak mieszkalnictwo, edukacja, praca z młodzieżą, indywidualne wsparcie oraz zdrowie i dobrostan.

Architektura federacyjna i kultura przejrzystości interpretacji

Ten przypadek stanowi istotny problem federacyjnego uczenia organizacyjnego. Lokalna autonomia pozwala dostosowywać działania do konkretnego kontekstu, jednak rozproszenie danych utrudnia wyciąganie wniosków na poziomie całej sieci. Z kolei centralizacja zwiększa porównywalność i ułatwia identyfikację wzorców, ale niesie ryzyko utraty lokalnych znaczeń oraz narzucenia jednolitej interpretacji. Rozwiązaniem nie jest zatem pełna centralizacja ani decentralizacja, lecz architektura warstwowa. Lokalne jednostki powinny zachować dane niezbędne do interpretacji własnego kontekstu, podczas gdy poziom centralny potrzebuje jedynie wspólnego rdzenia semantycznego, który umożliwi porównania. Takie podejście realizuje ideę interoperacyjności zamiast uniformizacji. Organizacja ucząca się jako federacja nie wymaga identyczności danych, lecz zdolności do rozpoznawania, czy podobne rezultaty powstają z podobnych mechanizmów, czy też te same działania przynoszą różne skutki w odmiennych warunkach.

Live Data może radykalnie skrócić czas między porażką a jej zauważeniem, jednak to kultura organizacyjna rozstrzyga, czy skróci się czas między zauważeniem a przyznaniem, że porażka

zaszła. To rozróżnienie jest kluczowe: technologia może zmniejszyć information latency, lecz nie znosi automatycznie psychological latency. Człowiek i organizacja potrzebują czasu, aby zaakceptować informację sprzeczną z ich tożsamością, reputacją czy interesem. Właśnie dlatego bezpieczeństwo psychologiczne – o którym autor pisze szerzej w kolejnej części – jest elementem infrastruktury danych, a nie tylko kwestią HR. Jeśli pracownik obawia się ujawnić niekorzystny wynik, dane zostaną opóźnione, przefiltrowane lub przeinterpretowane. Gdy fundator oczekuje permanentnego sukcesu, grantobiorca uczy się raportować przede wszystkim to, co zostanie uznane za sukces. W takim scenariuszu dashboard może być aktualizowany co godzinę, lecz system pozostaje poznawczo ślepy.

Zarządzanie wpływem wymaga zatem czegoś więcej niż transparentności zewnętrznej; potrzebuje transparency of interpretation – możliwości pokazania nie tylko samych danych, ale i sposobu ich rozumienia, rozważanych alternatyw oraz momentów zmiany zdania. Dwie organizacje mogą osiągać identyczne wyniki, dysponując zupełnie różnymi zdolnościami uczenia. Pierwsza natychmiast dostosuje narrację tak, by wynik pasował do wcześniejszych założeń. Druga uzna rozbieżność za okazję do rewizji Theory of Change.

Na tym poziomie pojawia się problem rozdzielenia learning i accountability. Dane służące uczeniu powinny zachęcać do eksperymentowania, wykrywania błędów i ujawniania niepowodzeń. Te same dane wykorzystywane do sankcjonowania mogą skłaniać do nadmiernej ostrożności i ukrywania problemów. Jeśli jeden wskaźnik służy jednocześnie obu celom, powstaje konflikt bodźców: im większe ryzyko utraty środków po ujawnieniu porażki, tym mniejsza motywacja do autentycznej eksploracji. Dlatego fundatorzy pragnący organizacji uczących się muszą zaakceptować koszt prawdziwego uczenia. Nie można jednocześnie wymagać eksperymentowania i penalizować każdego nieudanego doświadczenia. OECD w analizach innowacji sektora publicznego zauważa podobny problem: adaptacyjne uczenie i eksperymentowanie często pozostają zamknięte w laboratoriach innowacji, zamiast przenikać do rdzenia instytucji, a wyzwaniem jest właśnie stworzenie warunków dla double-loop learning.

Z tego względu jakość systemu Live Data należy oceniać nie po liczbie dashboardów, lecz po liczbie rzeczywistych decyzji zmienionych przez dane. To test znacznie bardziej wymagający. Organizacja może wykazać, że posiada platformę BI, ale czy potrafi wskazać program, który dzięki danym został zmodyfikowany? Czy wskaże miernik, którego przestała używać po odkryciu jego słabej trafności? Czy zamknęła działanie, które nie przynosiło skutku, lub zwiększyła finansowanie tam, gdzie outcome był najwyższy?

Od organizacji data-driven do epistemologii pokory

Czy po sygnale nierówności zmieniła zasady rekrutacji? Dopiero takie przypadki świadczą, że informacja przeszła pełną drogę od obserwacji do decyzji. Określenie „organizacja data-driven" jest w tym kontekście zbyt słabe. Dane nie powinny prowadzić organizacji mechanicznie, gdyż same w sobie wymagają interpretacji.

Bardziej adekwatna jest kategoria **evidence-responsive organisation**: organizacji, która potrafi aktualizować swoje przekonania i działania proporcjonalnie do jakości dowodów. Takie ujęcie łączy dyscyplinę naukową z odpowiedzialnością przywództwa. Zarząd nie może schować się za stwierdzeniem „tak pokazały dane", ponieważ to on wcześniej zdecydował, co mierzyć, jak interpretować i jakie konsekwencje uznać za istotne. Dojrzałość pojawia się dopiero na poziomie **double-loop governance**, kiedy dane o rezultatach nie tylko regulują wykonanie, lecz mogą prowadzić do rewizji celu, budżetu, portfela programów, struktury organizacyjnej i relacji z interesariuszami. Jeśli okazuje się, że organizacja od lat poprawia określony outcome, lecz coraz mniej odpowiada on rzeczywistym potrzebom ludzi, sukces wskaźnikowy powinien wywołać zmianę strategii. Jeśli Theory of Change okazała się błędna, nie należy jedynie zmieniać wartości docelowej – trzeba zmienić teorię.

Jeszcze wyższą formą dojrzałości jest **uczenie trzeciego porządku**, rozumiane jako zdolność do refleksji nad tym, jak organizacja tworzy wiedzę, komu ufa, jakie rodzaje dowodów uważa za wiarygodne i które głosy systematycznie wyklucza. To prowadzi zarządzanie wpływem w stronę epistemologii instytucjonalnej.

Organizacja może odkryć, że posiada sprawny system liczbowy, ale nie potrafi słyszeć doświadczenia beneficjentów; że ufa analizie centralnej bardziej niż wiedzy pracowników pierwszej linii; że premiuje dowody ilościowe nawet wtedy, gdy problem wymaga zrozumienia mechanizmu; albo że niepowodzenia są raportowane technicznie, lecz nie mają konsekwencji strategicznych. Taki sposób myślenia pozwala również inaczej zrozumieć różnicę między raportem rocznym a dashboardem. Nie chodzi o to, że pierwszy jest „stary", a drugi „nowoczesny". Raport może być znakomitym narzędziem refleksji długookresowej, porównania lat i publicznej odpowiedzialności, podczas gdy dashboard służy sterowaniu operacyjnemu.

Patologia powstaje wtedy, gdy jeden próbuje zastąpić drugi. **Organizacja potrzebuje kilku skal czasu poznawczego**: bieżącego monitoringu, kwartalnej interpretacji, rocznej ewaluacji oraz okresowej rewizji strategii i Theory of Change. To właśnie rozdzielenie skal pozwala uniknąć zarówno bezwładu, jak i hiperreaktywności. Dane krótkookresowe odpowiadają na pytanie, czy proces zachowuje się zgodnie z oczekiwaniem. Analiza średniookresowa pozwala oceniać

mechanizmy. Ewaluacja długookresowa bada trwałość i wpływ. Strategiczna rewizja może zakwestionować sam model.

Dojrzały system nie wybiera między nimi, lecz **układa je w hierarchię pętli uczenia**. Z tego punktu widzenia właściwa architektura wpływu mogłaby zostać opisana jako układ czterech sprzężeń. Pierwsze łączy działanie z natychmiastową informacją operacyjną. Drugie łączy outcomes z korektą programu. Trzecie łączy ewaluację z rewizją Theory of Change. Czwarte łączy doświadczenia wielu cykli z refleksją nad samym sposobem, w jaki organizacja formułuje cele i tworzy wiedzę. Im wyższa pętla, tym rzadsza powinna być jej aktywacja, ale tym głębsze mogą być jej konsekwencje.

Właśnie dlatego organizacja „Difference" nie jest organizacją posiadającą więcej informacji. Jest organizacją posiadającą **krótszą i uczciwszą drogę od informacji do korekty**. Jej przewaga nie polega na dashboardzie, lecz na zdolności powiedzenia: „nasza hipoteza była błędna", „ten program nie działa dla tej grupy", „ten rezultat przestał być wystarczającą definicją sukcesu", „musimy zmienić sposób działania" albo nawet „nasza dotychczasowa rola nie jest już najlepszą odpowiedzią na problem". Wtedy Live Data przestaje oznaczać technologiczną fascynację aktualnością danych i staje się elementem **organizacyjnej epistemologii pokory**. System ma nie tylko wiedzieć szybciej, ale szybciej odkrywać, gdzie wiedział źle. OECD coraz wyraźniej formułuje podobny postulat wobec polityk publicznych: monitoring, ewaluacja i feedback powinny łączyć outcomes z kolejnymi decyzjami, a w zmiennym otoczeniu służyć real-time learning zamiast wyłącznie ocenie ex post.

W tym punkcie praktyczny program Stevensona-Dodda spotyka się z cybernetyką, teorią organizacji i współczesną adaptacyjną governance. Tekst przechodzi zatem od mechanizmu uczenia do jego najtrudniejszego warunku kulturowego: **organizacyjnej zdolności do przyjęcia porażki jako informacji**. Sama pętla sprzężenia zwrotnego nie działa, jeżeli system usuwa sygnały niewygodne. Dlatego kolejnym krokiem będzie analiza **psychological safety**, kultury błędu, modeli „The Good, the Bad, the Ugly", Street League, **fail fast**, principal-agent, reputacyjnego ryzyka fundatorów oraz granicy między porażką generującą wiedzę a zaniedbaniem, które żadnej pochwały nie zasługuje.

> Ostatecznie najwyższą formą dojrzałości organizacji nie jest posiadanie najdokładniejszego dashboardu, lecz wykształcenie epistemologii pokory. Prawdziwy sukces mierzony jest bowiem nie liczbą potwierdzonych hipotez, lecz szybkością, z jaką system potrafi odkryć, gdzie wiedział źle. Czy jesteśmy zatem gotowi zamienić iluzję

> kontroli nad danymi na odwagę zmiany własnej tożsamości w odpowiedzi na rzeczywiste doświadczenia ludzi?

Inspiracje

[1] 

"The Difference How to Understand Measure and Maximise Social Impact" Matt Stevenson-Dodd

Practical Inspiration Publishing | ISBN: 9781788609562

Matt Stevenson-Dodd jest brytyjskim przedsiębiorcą społecznym, dyrektorem organizacji non-profit i konsultantem specjalizującym się w pomiarach wpływu społecznego. Przez dziewięć lat (2010–2019) pełnił funkcję dyrektora generalnego Street League, brytyjskiej organizacji charytatywnej wykorzystującej sport do walki z bezrobociem wśród młodzieży. Zasłynął tam jako pionier w zakresie transparentnego raportowania wpływu i otwartego reagowania na niepowodzenia organizacji. W 2019 roku założył Trust Impact, firmę konsultingową, której celem jest pomoc organizacjom charytatywnym, przedsiębiorstwom społecznym i organizacjom sektora publicznego w upraszczaniu, mierzeniu i maksymalizacji ich wpływu społecznego. Stevenson-Dodd pełnił również funkcję przewodniczącego Football Beyond Borders i Sport for Development Coalition, a także członka zarządu St John Ambulance, za co został przyjęty do grona Najczcigodniejszych Zakonu Świętego Jana w 2020 roku.

Matt Stevenson-Dodd *is a British social entrepreneur, non-profit executive, and consultant specializing in social impact measurement. He served for nine years as the Chief Executive Officer of Street League (2010–2019), a UK charity using sport to tackle youth unemployment, where he became well known for pioneering transparent impact reporting and openly addressing organizational failures. In 2019, he founded Trust Impact, a consultancy dedicated to helping charities, social enterprises, and public sector organizations simplify, measure, and maximize their social impact. Stevenson-Dodd has also served as Chair of Football Beyond Borders and the Sport for Development Coalition, as well as a trustee for St John Ambulance, for which he was invested into the Most Venerable Order of Saint John in 2020.*

Trust Impact

Literatura uzupełniająca

Książki

Książki autorów przywoływanych i literatura pokrewna:



[1] **"Inner Contradictions of Rigorous Research"**

Chris Argyris

1980

Artykuły naukowe

Prace autorów przywoływanych w tekście:

[1] **"Integrating the Individual and the Organization."**

Robert Henry Guest, Chris Argyris

1965 | American Sociological Review | DOI: 10.2307/2091786

Źródło

[2] **"Single-Loop and Double-Loop Models in Research on Decision Making"**

Chris Argyris

1976 | Administrative Science Quarterly | DOI: 10.2307/2391848

Źródło

[3] **"Overcoming Organizational Defenses: Facilitating Organizational Learning"**

Chris Argyris

1990

Źródło

[4] **"Organizational Learning II"**

Chris Argyris

1996

Źródło

[5] **"Organizational Learning: A Theory of Action Perspective"**

Ch. Argyris, Donald A. Schön

1997 | Revista Española de Investigaciones Sociológicas | DOI: 10.2307/40183951

Źródło

[6] **"Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions"**

Donald A. Schön

1987

Źródło

[7] "Book Review: Human Resources, Personnel, and Organizational Behavior: Organizational Learning II: Theory, Method, and Practice"

Davydd J. Greenwood, Chris Argyris, Donald A. Schön

1997 | Industrial and Labor Relations Review | DOI: 10.1177/001979399705000420

Źródło

[8] "Educating the Reflective Practitioner. Toward a New Design for Teaching and Learning in the Professions. The Jossey-Bass Higher Education Series."

Donald A. Schön

1987

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: 697db5cf-e030-41c7-ad4b-ada40bf50d13