Wszyscy chcą dobrze, a wychodzi jak zwykle

O strukturalnym rozmyciu odpowiedzialności i o otwartej ranie etycznej w zespołach budujących systemy AI.

Wojciech Bednaruk

od ponad 20 lat zajmuje się technologiami edukacyjnymi. Pracował nad wdrażaniem systemów zarządzania rozwojem w Kanadzie, Europie Środkowo-Wschodniej i Afryce, tworzył cyfrowe programy szkoleniowe. Wykładowca etyki sztucznej inteligencji na Polsko-Japońskiej Akademii Technik Komputerowych, członek Sekcji AWSI przy PTI.

W styczniu 2024 r. Mark Zuckerberg stanął przed Komisją Senatu Stanów Zjednoczonych. W sali siedzieli również rodzice, których dzieci okaleczyły się lub popełniły samobójstwa w wyniku korzystania z Instagrama i Facebooka. Wewnętrzne raporty Mety, ujawnione przez sygnalistkę Frances Haugen w 2021 r., dokumentują, że firma wiedziała o szkodliwym wpływie swoich platform na zdrowie psychiczne nastolatków i mimo to kontynuowała optymalizację algorytmów w celu pogłębiania uzależnienia.

Język biznesu

Jeden z senatorów zmusza Zuckerberga do przeproszenia rodzin ofiar. Zuckerberg odwraca się w stronę rodziców i mówi: „Przepraszam za wszystko, przez co przeszliście”. A potem wraca do pytań senatorów i powtarza jak zaklęcie: „Zadaniem Mety jest budować narzędzia wiodące w branży i wzmacniać pozycję rodziców”.

Ta wypowiedź nie jest kłamstwem, to całkowite zastąpienie języka odpowiedzialności językiem misji korporacyjnej. Rodzice pytają o swoje skrzywdzone dzieci. Zuckerberg odpowiada, że takie oto są nasze cele produktowe. Co musi się stać z człowiekiem i z organizacją, żeby w takim momencie, publicznie, przed rodzinami skrzywdzonych dzieci, podkreślać jakość własnego produktu? Odpowiedź na to pytanie daje historia biznesu i reguł nim rządzących.

W 1994 r. siedmiu prezesów największych firm tytoniowych stanęło przed Komisją Zdrowia Izby Reprezentantów Stanów Zjednoczonych. Każdy z nich, pod przysięgą,

złożył to samo oświadczenie: nikotyna nie uzależnia. Wewnętrzne dokumenty firm, ujawnione przez sygnalistów i w postępowaniach sądowych, dowodziły, że firmy wiedziały o uzależniającym charakterze nikotyny od lat 50. XX w. i nadal prowadziły badania nad zwiększeniem jej stężenia w ich produktach.

Prezesi nie kłamali z powodu osobistej deprawacji. Kłamali, bo przez dekady uczestniczyli w systemie, który wytworzył własny język, własne kryteria oceny i własną definicję odpowiedzialności zawodowej. Definicję, z której krzywda konsumentów została strukturalnie wykluczona.

Firma Ford Motor Company wiedziała, że zbiornik paliwa w modelu Pinto jest wadliwy i podatny na eksplozję przy tylnym zderzeniu. Ocenia się, że latach 70. XX w. w wyniku pożarów samochodów zginęło co najmniej 27 osób. Wewnętrzna analiza firmy wykazała, że koszt naprawy usterki we wszystkich wyprodukowanych egzemplarzach przekroczy szacowane koszty odszkodowań za śmierć i obrażenia. Ford zdecydował się nie naprawiać wady.

Decyzja ta nie była wyrazem złej woli jednej osoby. Była produktem procesu, w którym ludzkie życie stało się zmienną w równaniu optymalizacyjnym i nikt – na żadnym etapie – nie poczuł się odpowiedzialny za całość.

Philip Morris przygotował w 2001 r. dla rządu Czeskiej Republiki analizę, rzeczowo argumentując, że masowe palenie papierosów przyniesie oszczędności dla systemu opieki zdrowotnej ze względu na skrócenie życia palaczy i redukcję kosztów emerytalnych oraz opieki geriatrycznej. Autorzy dokumentu byli analitykami biznesowymi. Dobrze wykonywali swoją pracę. Prawdopodobnie nikt z nich nie uważał się za człowieka, który sporządza rachunek zysków, w której główną zmienną jest ludzka śmierć.

> *Te cztery przypadki łączy jeden mechanizm, czyli rozmycie odpowiedzialności moralnej przez podział pracy, hierarchię i język profesjonalizmu. Ten mechanizm działa dziś w zespołach projektujących systemy sztucznej inteligencji.*

Etyczne zawieszenie to nie anomalia

Od pięciu lat prowadzę zajęcia z etyki AI ze studentami, którzy budują systemy sztucznej inteligencji lub wchodzą na ten rynek. Co semestr obserwuję ten sam wzorzec zachowań. Studenci opisują sytuacje, w których czuli, że coś jest nie tak, a mimo to pracowali dalej. Powtarzają się trzy narracje:

Narracja 1: „Nie wiedziałem, że to będzie problematyczne".

Narracja 2: „Nie mogłem nic zrobić, nie miałem władzy, nie miałem narzędzi".

Narracja 3: „Byłem sam. Nikt inny tego nie widział, albo wszyscy widzieli, ale milczeli".

Każda z tych narracji jest półprawdą, która chroni system, a nie człowieka, który w nim pracuje. Takimi narracjami posługują się nadal pracownicy Forda, analitycy Philip Morrisa, prawnicy firm tytoniowych, inżynierowie Mety i moi studenci.

Ale jest coś, o czym te narracje milczą. To cena, którą płaci człowiek, który z nich korzysta. Nie mam tu na myśli kary zewnętrznej, wyroku sądowego, utraty pracy czy nieprzychylnych nagłówków na portalach informacyjnych. Mam na myśli coś, co obserwuję w pracy z ludźmi, którzy przez lata funkcjonowali w systemach, które ich niepokoiły. Chodzi o niegojącą się ranę etyczną.

Teza, którą stawiam, brzmi następująco: szkodliwe systemy AI nie powstają przez złą wolę. Powstają przez strukturalną niemożliwość przypisania odpowiedzialności moralnej w systemie, który ją celowo rozprasza między role i procesy. I każdy człowiek, który przez długi czas działa wbrew własnym wartościom w takim systemie, płaci za to cenę psychologiczną, której żadna narracja przetrwania nie eliminuje. Tylko ją odracza.

Jak powstaje krzywda

W 2012 r. zespół Facebooka przeprowadził eksperyment na 689 tys. użytkowników, manipulując treścią ich feedu w celu zbadania zjawiska zarażenia emocjonalnego. Miesiące przygotowań, spotkań projektowych, przeglądów kodu, recenzji i przez cały ten czas nikt nie zadał pytań: „Czy mamy zgodę tych ludzi?" lub „Czy nie wyrządzamy im krzywdy?". Nie dlatego, że członkowie zespołu byli nieetyczni. Dlatego, że projekt operował językiem badań naukowych, a zespół posługiwał się konceptami *istotności statystycznej*, a nie *krzywdy psychologicznej*.

To jest pułapka ramowania. Kiedy projekt zostaje przedstawiony jako wyzwanie techniczne, w umysłach członków zespołu uruchamia się zestaw koncepcyjnych narzędzi technicznych. Przestają widzieć człowieka po drugiej stronie.

Zuckerberg przed Senatem mówił językiem produktowym, bo przez lata organizacja uczyła go, że to jest właściwy ję-

zyk dla każdej sytuacji. „Zadaniem Mety jest budować narzędzia wiodące w branży". To *dictum* jest prawdziwe jako opis celów korporacyjnych, ale nie odpowiada na pytanie, kto za te cele płaci zdrowiem a czasem i życiem.

> *I właśnie ta ślepota, wyuczona, strukturalna, nagradzana, jest tym, co Albert Bandura opisał jako moralne rozłączenie: psychologiczny mechanizm, przez który człowiek oddziela swoje działania od ich moralnego znaczenia, żeby móc funkcjonować.*

Wspierającym mechanizmem społecznym w takiej sytuacji jest rozproszenie odpowiedzialności. Na codziennym spotkaniu projektowym padają pytania: czy to działa?, czy to jest szybkie?, czy to się skaluje? Nie pada pytanie: kto może ucierpieć przez to, co budujemy? Inżynier AI myśli: „To nie moja rola". Kierownik produktu myśli: „Programiści znają szczegóły techniczne". Wszyscy razem budują system AI zanurzeni w etycznej ślepocie. Każdy wykonuje swoją pracę dobrze. Krzywda jest efektem systemu jako całości i nie pojawia się w żadnym module, nie pojawia się w żadnym logu, nie jest zaadresowana w żadnym tickecie.

Kiedy przedstawiam studentom dylemat etyczny i pytam: „Czy widzicie tu problem etyczny?", zapada cisza. Wszyscy patrzą w laptopy. Po chwili ktoś powie niepewnie: „Myślę, że..." i nagle widzę pięć rąk w górze: „Tak, też to widziałem!". Wszyscy widzieli. Każdy myślał, że jest sam. To jest pluralistyczna ignorancja. Większość prywatnie odrzuca normę, ale publicznie ją akceptuje, bo każdy sądzi, że jest w mniejszości. Prezesi firm tytoniowych pod przysięgą byli produktem organizacji, która przez dekady wytwarzała dokładnie ten mechanizm. Każdy wiedział i każdy sądził, że pozostaje sam z tą wiedzą.

Rana, która nie goi się sama

Jest coś, o czym literatura etyki organizacyjnej mówi rzadziej niż o mechanizmach systemowych. Co dzieje się z konkretnym człowiekiem, który przez miesiące i lata funkcjonuje wbrew własnym wartościom?

Nie ma tu jednej odpowiedzi. Obserwuję trzy wzorce, które wynikają z rozmów z ludźmi z branży technologicznej i które rozpoznaję w historycznych przypadkach korporacyjnych.

Wzorzec pierwszy: znieczulenie
Część osób radzi sobie z raną etyczną, wypracowując dystans do własnych reakcji moralnych. Przestają zauważać niepokój. Nie dlatego, że go nie odczuwają, ale dlatego, że nauczyli się ignorować ranę tak sprawnie, że znika z pola uwagi. To jest mechanizm powtarzalnego działania wbrew własnym wartościom, bez zewnętrznych konsekwencji, co prowadzi do obniżenia progu wrażliwości. Człowiek, który przez lata pracował w firmie tytoniowej nad zwiększeniem skuteczności uzależniania, nie stracił kompasu moralnego z dnia na dzień. Tracił go przez tysiące małych momentów, w których nie zauważył, że coś go niepokoi.

Wzorzec drugi: racjonalizacja
Inni zachowują świadomość niepokoju, ale stosują narracyjny opatrunek: „Taki już jest świat". „Gdybym ja tego nie zrobił, zrobiłby ktoś inny". „Przynajmniej my robimy to lepiej niż konkurencja". Te narracje są psychologicznie funkcjonalne, pozwalają spać w nocy i utrzymywać spójność postrzegania siebie jako przyzwoitego człowieka. Są też precyzyjnie opisane w literaturze jako mechanizmy moralnego rozłączenia – zniekształcenia poznawcze, które pozwalają nam działać wbrew własnym zasadom bez poczucia winy. Analitycy Philip Morris, którzy obliczali zyski z ludzkiej śmierci, musieli stosować jakiś wariant tej narracji. Inaczej dokument po prostu by nie powstał.

Wzorzec trzeci: ból bez ujścia
Część ludzi nie znieczula się i nie racjonalizuje i przez to nosi ranę przez lata. Wiedzą, co zbudowali. Pamiętają moment, w którym mogli zadać pytanie, i nie zadali. Pamiętają projekt, przy którym milczeli, bo ocena wydajności nagradzała dostarczanie wyników, a nie czujność etyczną. Ten ból, jeśli nie znajduje przestrzeni do wypowiedzenia, prowadzi – jak wynika z rozmów z doświadczonymi pracownikami branży – do goryczy i rozczarowania. Nie wobec jednej firmy czy jednego projektu, ale wobec całej branży czy nawet całego świata. Wobec idei, że technologia AI może być czymś więcej niż narzędziem optymalizacji cudzych przychodów.

Badania nad wyuczoną bezradnością pokazują, jak powtarzane doświadczenia braku wpływu prowadzą do generalizowanej pasywności. Prowadzą do stanu, w którym inżynier AI przestaje podejmować próby działania nawet wtedy, gdy działanie jest możliwe. Człowiek przestaje wierzyć, że jego osąd moralny ma jakiekolwiek znaczenie w strukturze, w której pracuje. I ta utrata wiary jest być może kosztowniejsza niż każda konkretna decyzja projektowa, przy której milczał.

Obserwuję to u studentów, którzy wchodzą na rynek pracy z wrażliwością etyczną i po dwóch latach zaczynają posługiwać się z językiem, który tę wrażliwość zastąpił. „Moim zadaniem jest budowanie narzędzi wiodących w branży" – słyszałem to od prezesów. Teraz słyszę od studentów pierwszego roku, którzy jeszcze nie napisali pierwszej linijki produkcyjnego kodu. Nie kłamią. Naprawdę w to wierzą, bo tak branża definiuje profesjonalizm… To jest może najgroźniejszy mechanizm ze wszystkich – znieczulenie wyprzedzające ranę.

Trzy poziomy interwencji

Firmy tytoniowe miały działy prawne i CSR. Ford miał inżynierów bezpieczeństwa. Philip Morris miał analityków. Meta ma zespoły ds. bezpieczeństwa i odpowiedzialnej AI – przynajmniej do czasu, kiedy te zespoły zostały zwolnione. Obecność procedur nie zastępuje osądu moralnego. A polityki korporacyjne nie leczą ran, których istnienia i tak nie uznają.

Dlatego moje rekomendacje kieruję do szeregowych pracowników firm projektujących systemy AI. Nie do zarządów i liderów. Do programistów, analityków danych, inżynierów ML, projektantów UX, do ludzi, którzy codziennie podejmują dziesiątki małych decyzji, z których każda wydaje się techniczna, a żadna z osobna nie wygląda jak decyzja moralna. I do tych spośród nich, którzy noszą w sobie ranę etyczną.

Poziom pierwszy: **zachowajcie kontakt z własną percepcją moralną.**
Kiedy coś was niepokoi w projekcie, zróbcie notatkę. E-maill do siebie z minimalnym zestawem danych: data, projekt, obawa. Nie z powodu odpowiedzialności organizacyjnej, tylko dla zachowania kontaktu z własną percepcją. Za tydzień, za miesiąc, za rok zaczniecie wątpić: „Może przesadzałem?". Zapis odpowie, że nie.

To jest jednocześnie elementarna uczciwość wobec siebie i ochrona przed znieczuleniem. Pamięć moralna wymaga pielęgnacji, tak jak każda inna.

Poziom drugi: **przełamcie izolację, choćby raz.**
Izolacja nie jest przypadkiem, jest narzędziem systemu. Nagradzane jest dostarczanie wyników, nie czujność etyczna. Ktoś kiedyś zgłosił temat i został odsunięty od projektu. Wszyscy inni widzieli, co się stało. Stąd nauka – nie sprawiaj problemów. I zaczynacie wątpić we własną percepcję: „Może tylko ja to widzę? Może przesadzam?".

Ale często nie tylko wy macie wątpliwości. Dlatego zapytajcie jedną osobę – nie publicznie, nie na spotkaniu – ale w cztery oczy: „Czy ty też masz uwagi co do tego projektu?". Odkrycie, że nie jesteście sami, przynosi istotną zmianę – rana etyczna przestaje być odczuwana w izolacji. Przestanie się ona pogłębiać za sprawą samotności, w której dotąd była noszona.

Pracownicy Google'a, którzy w 2018 r. protestowali przeciwko kontraktowi wojskowemu Projekt Maven, zebrali ponad cztery tysiące podpisów pod petycją. Zdobycie każdego podpisu było efektem jednej rozmowy w cztery oczy.

Poziom trzeci: **zadajcie jedno pytanie więcej, ale we właściwym momencie.**
Na przeglądzie kodu, na planowaniu sprintu, na spotkaniu przed wdrożeniem zapytajcie: „Czy testowaliśmy to na użytkownikach z niepełnosprawnościami?", „Kto może być poszkodowany, jeśli model pomyli się systematycznie dla określonej grupy demograficznej?", „Czy mimo tego, że model nie przetwarza danych osobowych, w jakiś sposób zmienia zachowania ludzi?"

To nie jest sabotaż. To jest staranność zawodowa i jednocześnie ochrona własnej integralności moralnej. Każde zadane pytanie jest małym aktem uczciwości wobec samego siebie. Nie zmieni architektury systemu. Ale zdecyduje o tym, czy patrząc w lustro za pięć lat, rozpoznamy siebie.

> *Budowanie takiego języka, pytanie po pytaniu, spotkanie po spotkaniu, to właśnie zadanie szeregowego pracownika firmy AI. I to jest jednocześnie jedyna dostępna profilaktyka przed raną etyczną.*

Zuckerberg przed Senatem, prezesi firm tytoniowych pod przysięgą, inżynierowie Forda przy arkuszach kalkulacyjnych, analitycy Philip Morris przy modelach finansowych – żaden z nich nie był moralnym potworem. Wszyscy byli profesjonalistami funkcjonującymi w systemach, które zoptymalizowały ich działanie, być może wbrew ich własnym wartościom i dostarczyły im narracji, dzięki którym mogli wmówić sobie, że wszystko jest w porządku.

Studenci, którzy dziś wchodzą na rynek AI, zasługują na coś więcej niż system, który znieczuli ich przed etyczną raną, zanim zdążą ją poczuć. Zasługują na środowisko, które daje im język do nazywania tego, co widzą i poczucie struktury, która sprawi, że zgłaszanie zastrzeżeń jest zawodowo możliwe, a nie samobójcze. Moje zadanie, jako edukatora, polega na tym, żeby budować właśnie takie środowisko.

Tekst jest rozwinięciem tematu, który Autor zaprezentował na konferencji „AI made in Poland" https://www.aimadeinpoland.com/