STUDIA SOCJOLOGICZNE 2021
1 (240), 89–115
ISSN 0039−3371, e-ISSN 2545–2770
DOI: 10.24425/sts.2021.136280

Karolina Sztandar-Sztanderska
Uniwersytet Warszawski
Michał Kotnarowski
Instytut Filozofii i Socjologii PAN
Marianna Zieleńska
Uniwersytet Warszawski

CZY ALGORYTMY WPROWADZAJĄ W BŁĄD? METAANALIZA ALGORYTMU PROFILOWANIA BEZROBOTNYCH STOSOWANEGO W POLSCE[1]

Decyzje w polityce społecznej podejmowane z użyciem algorytmów wpływają na jakość życia ludzi na świecie. Niedostępność algorytmów utrudnia ocenę ich wiarygodności. Nie wiadomo, czy modele statystyczne dobrano i zastosowano prawidłowo. Czy dane były wiarygodne? Autorzy podejmują ten ogólniejszy problem na przykładzie jednego z pierwszych algorytmów wdrożonych w Polsce: narzędzia profilowania bezrobotnych. Algorytm miał mierzyć potencjał osób bezrobotnych i na tej podstawie pomóc dzielić je na grupy o zróżnicowanym prawie dostępu do aktywizacji zawodowej. Opierając się na analizie dokumentów urzędowych, uzupełnionych o dane jakościowe i ilościowe, autorzy prześledzili decyzje podejmowane podczas konstrukcji algorytmu i dokonali metaanalizy

Karolina Sztandar-Sztanderska, Wydział Socjologii UW, e-mail: k.sztanderska@gmail.com, ORCID 0000-0001-7292-0728; Michał Kotnarowski, Instytut Filozofii i Socjologii PAN, e-mail: kotnarowski@gmail.com, ORCID 0000-0002-3468-0732; Marianna Zieleńska, Wydział Socjologii UW, e-mail: zielenskam@is.uw.edu.pl; ORCID 0000-0001-6297-2671.

Źródło finansowania: Projekt „Technologie informacyjne w polityce publicznej. Krytyczna analiza profilowania bezrobotnych w Polsce" finansowany przez Narodowe Centrum Nauki w ramach programu OPUS (2016/23/B/HS5/00889).

[1] Artykuł powstał dzięki współpracy z Fundacją Panoptykon, która wygrała dostęp do algorytmu profilowania w sądzie. Tekst przygotowaliśmy w ramach projektu *Technologie informacyjne w polityce publicznej. Krytyczna analiza profilowania bezrobotnych w Polsce* finansowanego przez Narodowe Centrum Nauki (2016/23/B/HS5/00889), kierowanego przez K. Sztandar-Sztanderską. Dziękujemy anonimowemu Recenzentowi za uwagi do wstępnej wersji artykułu. Za cenne sugestie odnośnie wcześniejszych wersji tekstu dziękujemy Alicji Pałęckiej Annie Kiersztyn, Wiesławie Kozek, Małgorzacie Sikorskiej, Joannie Mazur, Jędrzejowi Niklasowi, Barbarze Godlewskiej-Bujok. Alicji Pałęckiej dziękujemy również za koordynację badania i uporządkowanie danych, a firmie Dyspersja za profesjonalną realizację badania ilościowego. W tekście wykorzystujemy też wyniki badań pilotażowych: 1) Ekspertyzę dla Polskiego Komitetu EAPN ; Projekty 2) *Zmierzyć, zważyć, policzyć, zaklasyfikować* (DSM 112900/16); 3) *„Wyprofilować" bezrobotnego* (DSM 110400/66, DSM 110 400/72), 4) *Profiling the Unemployed in Poland. Social and Political Implications of Algorithmic Decision Making*, realizowany na zlecenie Fundacji Panoptykon. Dziękujemy też dyrekcji i pracownikom Departamentu Rynku Pracy z Ministerstwa Rodziny Pracy i Polityki Społecznej (MRPiPS), że zdecydowali się przekazać nam dalsze informacje nt. profilowania, widząc sens badania tego algorytmu.

statystycznej tego narzędzia. W artykule dowodzą, że algorytm profilowania nie spełniał podstawowych standardów metodologicznych: dane o osobach bezrobotnych były nierzetelne, błędnie zastosowano model psychometryczny, nieprawidłowo skonceptualizowano podstawową zmienną, formuły matematycznej nie dostosowywano do wyników analiz, lecz do poczynionych z góry założeń.

Słowa kluczowe: metodologia; algorytmy; polityka społeczna; profilowanie bezrobotnych; zautomatyzowane podejmowanie decyzji

Karolina Sztandar-Sztanderska, University of Warsaw, Faculty of Sociology
Michał Kotnarowski, Institute of Philosophy and Sociology, Polish Academy of Sciences
Marianna E. Zieleńska, University of Warsaw, Faculty of Sociology

Are the algorithms misleading? Meta-analysis of the algorithm for profiling of the unemployed used in Poland

Social policy decisions based on algorithms affect the quality of life of people. Yet, the access to algorithms is restricted, which makes it difficult to assess their credibility. In result, it often remains unknown whether the statistical models were correctly applied or based on reliable data. The authors address this more general problem, by referring to the example of the tool for profiling the unemployed implemented in Poland. The profiling algorithm was to measure the potential of the unemployed in order to divide them into groups with different access to activation measures. Based on the analysis of the official documents, supplemented with qualitative and quantitative data, the authors performed a statistical meta-analysis of this tool. They prove that the profiling algorithm did not meet the basic methodological standards in terms of data quality and the selection and application of the statistical model.

Key words: methodology; algorithms; social policy; profiling of the unemployed; automated decision making

Wstęp

W artykule podejmujemy problem wiarygodności diagnoz generowanych przez algorytmy, odwołując się do przykładu profilowania bezrobotnych. Analizujemy instrument wykorzystywany przez powiatowe urzędy pracy (PUP)[2] w latach 2014–2019. Algorytm miał mierzyć potencjał każdej osoby bezrobotnej, by dopasować pomoc do indywidualnych potrzeb i doprowadzić do zatrudnienia. W zależności od wyniku pomiaru program sugerował urzędnikowi przypisanie osoby bezrobotnej do jednej z trzech kategorii. Ta napędzana matematyką klasyfikacja służyła do „sortowania" ludzi (Lyon 2005; Bowker, Star 2000) na grupy o zróżnicowanym zakresie praw i obowiązków. Miała więc wpływ na kształt

[2] Używając skrótu PUP odnosimy się do powiatowych i miejskich instytucji.

„obywatelstwa społecznego" – by przywołać pojęcie ukute przez brytyjskiego socjologa Thomasa Marshalla (1950)[3].

Modele matematyczne zintegrowane z oprogramowaniem stosuje się na niespotykaną dotąd skalę, a ich znaczenie rośnie (zob. Pasquale 2015; O'Neil 2017; Eubanks 2017). Istotne jest więc weryfikowanie jakości algorytmów i identyfikowanie błędów (zob. np. Zweig et al. 2018; Berman 2018; Mittelstadt et al. 2016; Rieke et al. 2018). Niestety zazwyczaj nie jest to możliwe, ponieważ dostęp do nich jest strzeżony, co uzasadnia się przeciwdziałaniem nadużyciom i tajemnicą handlową (Pasquale 2015: 4).

Narzędzie profilowania bezrobotnych zastosowane w Polsce jest pod tym względem wyjątkowe, gdyż udało się otworzyć „czarną skrzynkę" technologii. Mieliśmy do czynienia z rzadką sytuacją, gdy „sygnalista wszczyna postępowanie sądowe lub ujawnia informacje" (Pasquale 2015: 4). Analizujemy dokumenty, które wyciekły z PUP do mediów oraz materiały statystyczne otrzymane od Fundacji Panoptykon po wygranej sprawie sądowej z Ministerstwem (Rodziny) Pracy i Polityki Społecznej (M(R)PiPS) o ujawnienie algorytmu[4]. Nasz wkład zasadza się na tym materiale, uzupełnionym o dalsze dane jakościowe i ilościowe.

Choć profilowanie wywołało liczne kontrowersje i zostało wycofane po niekorzystnym wyroku Trybunału Konstytucyjnego, wciąż nikt nie przeanalizował ani zastosowanego modelu statystycznego, ani wiarygodności przetwarzanych przezeń danych. Inni naukowcy, nie znając konstrukcji modelu przywoływali oficjalne uzasadnienia tego narzędzia (Wiśniewski, Wojdyło-Preisner 2015), pisali o technikach profilowania bezrobotnych stosowanych w innych krajach (Wojdyło-Preisner 2009) lub sprawdzali, co o profilowaniu sądzi kadra PUP (Herman-Pawłowska i in. 2016). My z kolei, dotychczas skupialiśmy się na innych jego wymiarach. Pokazywaliśmy, jak technologia może niepostrzeżenie zmienić proces podejmowania decyzji w obszarze bezrobocia (Sztandar-Sztanderska, Zieleńska 2018), analizowaliśmy stojący za nią ideał normatywny (Sztandar-Sztanderska, Zieleńska 2020) oraz zwracaliśmy uwagę na brak dostępu do programów aktywizacji zawodowej dla osób zaliczanych do ostatniego profilu (Niklas i in. 2015: 32, 36–37; Sztandar-Sztanderska 2013: 27–28).

Artykuł składa się z pięciu części. W pierwszej uzasadniamy znaczenie projektu, odwołując się do badań o algorytmach w polityce społecznej. Następnie opisujemy metodologię badania. W trzeciej i czwartej części przedstawiamy

[3] Więcej na temat obywatelstwa społecznego zob. (Theiss 2018).

[4] Fundacja Panoptykon wystąpiła latem 2015 r. do MPiPS kierowanego przez Władysława Kosiniaka-Kamysza (PSL) z wnioskiem o udostępnienie algorytmu wpływającego na zakres pomocy udzielanej bezrobotnym jako informacji publicznej. Po odmowie, zaskarżyła ministerstwo w sądzie. Wyrok korzystny dla fundacji zapadł 05.04.2016, gdy resortem o nazwie Ministerstwo Rodziny Pracy i Polityki Społecznej (MRPIPS) kierowała Elżbieta Rafalska (PiS).

wyniki. Najpierw rekonstruujemy instytucjonalny kontekst profilowania. Następnie oceniamy jakość modelu pomiarowego, sprawdzając rzetelność danych używanych do profilowania oraz poprawność operacji statystycznych. W podsumowaniu odnosimy wyniki do ogólniejszego problemu stosowania algorytmów w polityce społecznej.

Algorytmy w polityce społecznej

Zinformatyzowane modele statystyczne stosuje się w różnych obszarach polityki społecznej. W edukacji używa się ich do oceny pracy nauczycieli (O'Neil 2017), w wymiarze sprawiedliwości do przewidywania prawdopodobieństwa recydywy osoby skarżonej i orzekania o wymiarze wyroku lub kaucji (Harcourt 2005), w systemie zabezpieczenia społecznego do typowania osób podejrzanych o wyłudzanie świadczeń (Dubois et al. 2018). W polityce rynku pracy wykorzystuje się je do decydowania, których bezrobotnych aktywizować za pomocą programów finansowanych ze środków publicznych (Allhutter et al. 2020; Corbanese, Rosas 2017; Desiere et al. 2019; Mozzana 2019; Wiśniewski, Wojdyło-Preisner 2013). Wiele decyzji wymagających kiedyś profesjonalnej oceny jest obecnie „z góry zaprogramowanych" (Bovens, Zouridis 2002: 175; zob. też Citron, Pasquale 2014). Z informacji ONZ wynika, że technologie informacyjno-komunikacyjne (ICT) stosowane w polityce społecznej wpływają na życie miliardów ludzi na świecie (Alston 2019). Dlatego też w agendach międzynarodowych i dyskursie publicznym mówi się o „cyfrowym welfare state" (Alston 2019).

Co istotne, w odróżnieniu od reguł prawnych, zasady działania ICT są niejawne. Nie poddaje się ich naukowej weryfikacji, konsultacjom społecznym czy publicznej dyskusji, tak jak to ma (lub powinno mieć) miejsce w przypadku uchwalania prawa (Bovens, Zouridis 2002; Zouridis et al. 2020). W efekcie, nikt nie wie, czy diagnozy dokonywane przez algorytm są prawidłowe; jakie decyzje kryją się w technologii; czy są one zgodne z prawem – np. przepisami antydyskryminacyjnymi (Mazur 2018) – oraz czy byłyby odbierane jako kontrowersyjne lub niesprawiedliwe (Mittelstadt et al. 2016; O'Neil 2017; Pasquale 2015). Ponadto projektanci ICT – anonimowi statystycy, informatycy, programiści oraz ich pracodawcy i zleceniodawcy – nie są w swoich wyborach neutralni i pozbawieni uprzedzeń (zob. np. Piwowar 2019; Criado-Perez 2019; Angwin et al. 2016). Rozstrzygnięcia zaprogramowanych przez nich systemów są często niepodważalne. Brakuje mechanizmów kontrolnych, które pozwalałyby na weryfikację decyzji podejmowanych przez projektantów, w odróżnieniu od decyzji polityków, które weryfikowane są przez proces wyborczy lub decyzji urzędników, które można zaskarżyć na podstawie przepisów prawa (Wedel 2014: 30 i następne; Zouridis et al. 2020). Podsumowując, fundamentalne dla jakości demokracji

wymogi przejrzystości (*transparency*) i rozliczalności (*accountability*) nie mają zastosowania wobec algorytmów i ich twórców.

Mimo to technologie przedstawia się jako sposób na oszczędności, efektywną, sprawiedliwą i pozbawioną uprzedzeń administrację. Obiektywnością i przeciwdziałaniem nadużyciom uzasadnia się zamknięcie modelu w czarnej skrzynce. Tymczasem wiedza wytwarzana przez algorytmy ma charakter probabilistyczny, a w wyniki bardzo często wpisany jest błąd pomiaru (Berman 2018; Mittelstadt et al. 2016). Ponadto, wnioski dotyczą współwystępowania zjawisk (korelacje), a nie związków przyczynowo-skutkowych. Z tych powodów algorytmiczne wyniki trudno uznać za wystarczającą przesłankę do działania (Mittelstadt et al. 2016: 4). Omylność czasem idzie też w parze ze stronniczością (*bias*): np. algorytm COMPAS generował wyniki błędnie dodatnie przy przewidywaniu, że czarnoskórzy Amerykanie powtórnie popełnią przestępstwo i błędnie ujemne w przypadku białych (Angwin et al. 2016). Badacze krytykują też decyzje kryjące się w algorytmach, pod kątem ich negatywnego wpływu na nierówności i wykluczenie (Eubanks 2017; O'Neil 2017; Mileszczyk et al. 2019). Warto też pamiętać, że ludziom obsługującym technologie trudno jest je nadzorować i korygować ich ewentualne błędy, gdy pracują pod presją czasu, brak im specjalistycznych kompetencji lub pozycji gwarantującej decyzyjność (Elish 2019).

Metodologia

W artykule wykorzystujemy materiał zbierany przez 7 lat. Gromadząc i analizując dane, stosowaliśmy zasady triangulacji badaczy i metod (Denzin 1970).

Triangulacja badaczy polegała na zaangażowaniu w projekt osób o zróżnicowanym zapleczu teoretycznym i metodologicznym, a także – co ważne w kontekście wcześniejszej współpracy współautorki tego artykułu z aktywistami Fundacji Panoptykon (Niklas et al. 2015) – o zróżnicowanych poglądach na zastosowanie algorytmów w polityce publicznej. Praca zespołowa zasadzała się na łączeniu różnych kompetencji: umiejętności analizy przepisów (Sztandar-Sztanderska 2013, 2016), znajomości realiów funkcjonowania ministerstwa inicjującego profilowanie (Zieleńska 2015) oraz wdrażających je PUP (Sztandar-Sztanderska 2013, 2016), wiedzy z zakresu statystyki, metodologii badań jakościowych, ilościowych i psychometrycznych, doświadczenia w prowadzeniu ewaluacji (Konarski, Kotnarowski 2007; Sztandar-Sztanderska 2010, 2013; Zieleńska, Tomasik 2010), zamiłowania do socjologii nauki i technologii (Niklas et al. 2015; Sztandar-Sztanderska, Zieleńska 2020).

Triangulacja metod polegała zaś na wykorzystaniu uzupełniających się danych zastanych, jakościowych, ilościowych, a dokładniej:

- Kwestionariusza profilowania, za pomocą którego zbierano i przetwarzano dane o bezrobotnych [MPiPS1][5].
- Raportów statystycznych z wynikami pilotażu kwestionariusza, opisem wariantów algorytmu oraz danych zastanych dokumentujących proces tworzenia i wdrażania narzędzia [MPiPS2, 3, 4, 5, 6, 7, 10, 11].
- Podręcznika profilowania zawierającego wskazówki dla kadry PUP jak profilować [MPiPS8].
- Aktów prawnych regulujących profilowanie [USTAWA, ROZPORZĄDZENIE].
- Wyników reprezentatywnego badania sondażowego przeprowadzonego w ramach naszego projektu w 2019 roku z kadrą PUP techniką CAWI na próbie 190 PUP[6] [CAWI].
- Pięciu wywiadów częściowo zestandaryzowanych z twórcami narzędzia i uczestnikami procesu legislacyjnego [WYWIAD1;2;3;4;5].
- Dwóch niejawnych obserwacji uczestniczących: jednej w komisji sejmowej, drugiej w MRPiPS [OBSERWACJA1; 2].

Opierając się na tych źródłach, przeanalizowaliśmy decyzje metodologiczne oraz dane używane do profilowania i konstrukcję algorytmu pod kątem podstawowych standardów naukowych. Do oceny wiarygodności algorytmicznych diagnoz dochodziliśmy stopniowo, weryfikując wnioski cząstkowe z materiałem badawczym, kodowanym w programie MAXQDA. Ponadto, zgodnie z zaleceniami literatury przedmiotu, zrekonstruowaliśmy szerszy kontekst, w ramach którego wdrożono to narzędzie (Peña Gangadharan, Niklas 2019), pokazując cele i konsekwencje reformy oraz naukową legitymizację profilowania.

Instytucjonalny kontekst profilowania – wyniki badania

Celem reformy wprowadzającej profilowanie było zwiększenie skuteczności pracy PUP [MPiPS9: 4]. Profilowanie było przedstawiane jako zestandaryzowany mechanizm dystrybucji środków. Przekonywano, że pomoże dopasować programy aktywnej polityki rynku pracy (ALMP) do indywidualnych potrzeb i sprawi, że bezrobotni podejmą pracę [MPiPS9: 4, 12]. Stworzenie algorytmu wyłaniającego grupy „homogeniczne" pod kątem potrzeb miało też ograniczyć „błąd

[5] W aneksie umieszczamy spis analizowanych dokumentów z kodami, które podajemy w nawiasach kwadratowych.

[6] Urzędy wylosowano w 4 warstwach wyróżnionych ze względu na stopę bezrobocia rejestrowanego. Następnie spośród wszystkich pracowników PUP realizujących profilowanie wylosowano jednego, z którym przeprowadzono wywiad. Treść ankiety uwzględniała wstępne wyniki badań jakościowych o profilowaniu zrealizowanych wcześniej w 4 powiatach w ramach naszego projektu.

subiektywizacji" [MPiPS8: 4]. Zdaniem MPiPS ryzyko tego błędu było wysokie, gdy kadra PUP bez żadnych matematycznych narzędzi decydowałaby, jakie ALMP przyznać danej osobie [MPiPS8: 4; OBSERWACJA2]. Profilowanie miało zagwarantować, że kryteria będą jednolite, a podział środków na ALMP będzie oparty na „racjonalnych przesłankach i mierzalnych wartościach" [MPiPS4: 4].

Nieoficjalnie, dowiedzieliśmy się też, że profilowanie było sposobem radzenia sobie z brakami kadrowymi w PUP i niewystarczającą pulą środków na ALMP [OBSERWACJA2]. Nie rozważano możliwości zwiększenia nakładów, tylko szukano rozwiązań, jak alokować ograniczone zasoby między wielu bezrobotnych, identyfikując osoby, o których sądzono, że nie chcą podjąć pracy, ponieważ rejestrują się w PUP, by uzyskać dostęp do ubezpieczenia zdrowotnego [MPiPS8: 4].

1. Zróżnicowany zakres praw i obowiązków

W ustawie wprowadzającej profilowanie zawarto rozróżnienie na trzy grupy bezrobotnych: nazywane I, II i III profilem pomocy. Grupą priorytetową był II profil pomocy, dla którego ustawodawca przewidział najwięcej narzędzi [USTAWA: art. 33 ust. 2c]. Większe prawa wiązały się też z większym zakresem obowiązków, gdyż odmowa współpracy z PUP groziła wyrejestrowaniem i utratą uprawnień. Osoby zaklasyfikowane do II profilu miały być objęte pośrednictwem pracy i poradnictwem zawodowym [USTAWA: art. 33 ust. 2c]. Dla tej kategorii zaplanowano też opcjonalnie inne programy, umożliwiające podniesienie kwalifikacji, zdobycie doświadczenia, podjęcie pracy subsydiowanej lub samozatrudnienia czy wsparcie mobilności [USTAWA: art. 33 ust. 2c].

Natomiast w przypadku I i III profilu pakiety ALMP były ograniczone. Osoby zaklasyfikowane do I profilu miały być przede wszystkim objęte pośrednictwem pracy [USTAWA: art. 33 ust. 2c]. W „uzasadnionych przypadkach", ustawodawca dopuścił też inne ALMP [USTAWA: art. 33 ust. 2c]. Pakiet ALMP był najbardziej okrojony dla III profilu. Dla tej grupy dedykowano najtańsze narzędzia lub działania fakultatywne. W efekcie osoby doń zaklasyfikowane nie miały prawa (ale też obowiązku) korzystać z większości ALMP realizowanych przez PUP. Programy dla III profilu nie były dostępne na terenie całego kraju, a tam gdzie były dostępne, obejmowały tylko niewielki odsetek bezrobotnych (Niklas et al. 2015; Herman-Pawłowska et al. 2016: 61, [CAWI]). Taką diagnozę sformułowaną w trzech niezależnych badaniach potwierdziła Najwyższa Izba Kontroli konkludując, że „brak wsparcia ze strony powiatowych urzędów pracy (…) przyczyniał się do utrwalenia bezrobocia" wśród tej grupy [NIK2: 47].

2. Kluczowe znaczenie decyzji ukrytych w czarnej skrzynce

Przepisy nie definiowały jednak, kogo i na podstawie jakich kryteriów algorytm zaklasyfikuje do poszczególnych profili. Odpowiedzi na te pytania

były ukryte w ICT. Algorytm przetwarzał dane o osobie bezrobotnej. Były one częściowo pobierane z bazy danych PUP, a w większości uzupełniane za pomocą elektronicznego kwestionariusza przez pracownika PUP w trakcie rozmowy z osobą bezrobotną. Następnie algorytm zliczał punkty przypisane każdej z zestandaryzowanych odpowiedzi oraz dokonywał wstępnej klasyfikacji. Co istotne, resort pracy odmawiał ujawnienia treści kwestionariusza, punktacji odpowiedzi oraz zasad podziału na profile. Odmowę uzasadniano dbałością o jakość pomiaru i przeciwdziałaniem nadużyciom [WSA, MPIPS10]. Przepisy nie przewidywały też możliwości zakwestionowania przez osobę bezrobotną wyniku profilowania (Godlewska-Bujok 2020)[7].

Możliwość korekty automatycznej klasyfikacji przysługiwała za to pracownikom PUP, pod warunkiem, że uzasadnili ją w systemie [MPiPS8]. Jednak stosunkowo rzadko z tej opcji korzystali [MPiPS4, CAWI]. W pierwszych miesiącach system działał niemal na zasadzie automatycznego podejmowania decyzji [MPiPS4][8]. Z badania ankietowego wynika zaś, że również pięć lat po wprowadzeniu profilowania korygowanie automatycznej klasyfikacji nie było powszechną praktyką [CAWI][9].

3. Naukowa legitymizacja

Odmawiając ujawnienia informacji o algorytmie, ministerstwo starało się jednocześnie przedstawić go jako naukowy [WSA, MPIPS10]. Widać to w języku, którego przedstawiciele resortu pracy używali, opisując profilowanie. Ustalenie profilu miało następować w wyniku badania dwóch „zmiennych", uwzględniających szereg „czynników" [MPiPS8: 6; MRPiPS1]. Do pozyskiwania danych o bezrobotnych służył „standaryzowany kwestionariusz", który wraz z algorytmem wypracowano po „pilotażu" i testach na próbie „6 605 osób bezrobotnych" [MPiPS8; MRPiPS1]. Ostatecznym potwierdzeniem wiarygodności algorytmicznych diagnoz miały być – zdaniem przedstawicieli ministerstwa – satysfakcjonujące wyniki analizy rzetelności skal metodą Alfa Cronbacha [MRPiPS1]. Ta statystyczna retoryka skłoniła nas do sprawdzenia, w jakim

[7] To rozwiązanie skrytykował Rzecznik Praw Obywatelskich [RPO].

[8] Z danych MPiPS wiadomo, że w pierwszym półroczu profilowania profil wygenerowany przez algorytm był zaakceptowany przez kadrę PUP w 99,4% przypadków [MPiPS4]. Ponad połowa profilujących (51,7%) nie dokonała ani jednej korekty. Żadnej korekty nie dokonano w prawie 1/3 PUP (32%) [ibid.]. Prawdopodobnie w późniejszych latach, korekty profilu zdarzały się częściej i system funkcjonował na zasadzie częściowej-automatyzacji.

[9] Wyniki CAWI z 2019 r. wskazują, że w sytuacji, gdy automatyczny wynik wydał się pracownikom PUP nietrafny, 9% ankietowanych nigdy go nie korygowało; 24% przeważnie go nie korygowało, 9% korygowało go w połowie przypadków [CAWI_2019]. Osoby, które zawsze korygowały profil, gdy wydał im się nietrafny, stanowiły 18%, a 40% robiło to przeważnie [ibid.].

stopniu było to poprawnie skonstruowane narzędzie i czy dane których używano do generowania wyników można uznać za wiarygodne.

Wyniki metaanalizy algorytmu profilowania

By ocenić jakość modelu pomiarowego, sięgnęliśmy do „kuchni statystycznej". Zrekonstruowaliśmy kolejne decyzje metodologiczne, a następnie odnieśliśmy je do standardów naukowych. Wyniki analizy przedstawimy w podziale na dwa poziomy, wyróżniane w literaturze o algorytmach: poziom danych wejściowych i poziom operacji statystycznych (Piwowar 2019; Mittelstadt et al. 2016). Dane wejściowe mają znaczenie, ponieważ błędne dane generują błędne wyniki (Babbie 2004). Jeśli zaś systematyczne błędy pojawią się już w zestawie danych użytych do konstrukcji algorytmu, wówczas skonstruowany algorytm też obciążony będzie błędem (Mittelstadt et al. 2016). Na wiarygodność algorytmicznych diagnoz wpływają również operacje statystyczne: między innymi wybór modelu pomiarowego czy poprawność jego zastosowania.

1. Niska wiarygodność danych wejściowych

Rekonstruując proces zbierania danych o bezrobotnych, zidentyfikowaliśmy 3 czynniki, które zaważyły na ich niskiej rzetelności. Po pierwsze, osoby bezrobotne miały ograniczoną swobodę wypowiedzi. Po drugie, kwestionariusz używany do profilowania był złej jakości. Po trzecie, nieprawidłowo zastosowano pytania otwarte prekodowane, czyli pytania, które choć brzmią jak pytania otwarte, mają zamkniętą kafeterię odpowiedzi, a wyboru zaznaczanej opcji dokonuje ankieter (Brill 2008) – w przypadku profilowania robi to pracownik PUP.

1.1 Ograniczona swoboda wypowiedzi

Na rzetelność danych o osobach bezrobotnych negatywnie wpływał specyficzny dla PUP układ ról urzędnik–bezrobotny daleki od dobrowolności właściwej dla relacji ankieter–respondent. Odmowa profilowania skutkowała utratą statusu bezrobotnego [USTAWA]. Przymusowość profilowania stawiała pod znakiem zapytania szczerość odpowiedzi, co zmniejszało wiarygodność danych. Ponadto osoby bezrobotne nie były anonimowe i zdawały sobie sprawę, że od wyrażanych opinii zależy ich ścieżka w PUP. Profilowanie przeprowadzał pracownik PUP (tzw. doradca klienta), który miał mobilizować bezrobotnego do podjęcia pracy i w przypadku nieprzestrzegania reguł karać go wyrejestrowaniem. W interesie profilowanych leżało więc wyrażanie poglądów zgodnych z obowiązującymi w PUP normami. Wzmacniało to efekt społecznych oczekiwań i efekt ankieterski, czyli skłonność do udzielania odpowiedzi, które jawią się respondentom jako pożądane i spełniają oczekiwania osób zadających

pytania (Sułek 2002: 55). Przejawy tych zjawisk odnotowano w raporcie z pilotażu narzędzia [MPiPS2:21–22, 25–26].

1.2. Niska jakość kwestionariusza

Rzetelność danych zależała bezpośrednio od jakości elektronicznego kwestionariusza, za pomocą którego doradcy klienta wprowadzali dane do systemu w trakcie rozmowy z osobą bezrobotną[10]. Według standardów badań ilościowych, pytania ankietowe i kafeterie odpowiedzi powinny być sformułowane w sposób prosty, zrozumiały, jednoznaczny, niewartościujący (Babbie 2003; Krosnick, Presser 2010; Sułek 2002).

Tymczasem część pytań była jawnie sugerująca: np. pytanie *Czy szuka lub szukał Pan/Pani samodzielnie pracy?* [MPiPS1]. Inne dotyczyły kwestii, na które profilowani nie znali odpowiedzi i rodziły ryzyko odpowiedzi artefaktualnych. Za przykład może posłużyć pytanie: *Jak Pan/Pani sądzi czy w najbliższym czasie samodzielnie znajdzie Pan/Pani pracę?* [MPiPS1]. Niektóre pytania były zaś zbyt ogólne i niejednoznaczne, co stoi w sprzeczności z wymogiem precyzji: respondent powinien wiedzieć „dokładnie, o co pyta badacz" (Babbie 2004: 271). Najwięcej trudności interpretacyjnych sprawiały pytania: *Proszę wskazać przyczyny utrudniające Pani/Panu podjęcie pracy?* oraz *Co jest Pan/Pani w stanie zrobić w celu zwiększenia swoich szans na podjęcie pracy?* [MPiPS1; 2] (zob. też: Herman-Pawłowska et al. 2016: 55). Doradcy klienta mieli je zadawać, nie pokazując zestawu możliwych odpowiedzi [MPiPS8], w efekcie bezrobotni nie rozumieli, jak rozbudowanej wypowiedzi się od nich oczekuje [MPiPS2].

Hipotezę, że pytania kwestionariuszowe były niezrozumiałe dla grupy docelowej, potwierdzają wyniki reprezentatywnego badania ankietowego w PUP [CAWI]: w 2019 roku aż 66% doradców klienta uznało, że osoby bezrobotne miały trudności z odpowiedzią na niektóre pytania zawarte w kwestionariuszu (15% ankietowanych zaznaczyło opcję „zdecydowanie tak", a 51% „raczej tak") (podobne wyniki otrzymano we wcześniejszym badaniu, zob. Herman-Pawłowska et al. 2016: 54–55). Ponadto, wśród ankietowanych nie znalazł się ani jeden pracownik, który nie pomagał bezrobotnym zrozumieć pytań [CAWI]. Aż 45% ankietowanych doradców klienta robiło to zawsze, a dalsze 53% robiło to w co najmniej połowie przypadków profilowania [CAWI].

[10] Do profilowania używano też ośmiu typu danych zaciąganych z systemu Syriusz^{Std}: wiek, płeć, wykształcenie, okres wcześniejszego doświadczenia zawodowego, znajomość języków obcych, stopień niepełnosprawności, czas pozostawania bez pracy, liczba odmów oferowanej przez PUP pomocy. Nie mamy podstaw empirycznych, by mówić o jakości tych danych, więc skupiamy się wyłącznie na danych o bezrobotnych wprowadzanych do systemu przez urzędników po wywiadzie w oparciu o kolejnych 16 pytań kwestionariusza.

1.3. Niewłaściwe zastosowanie pytań otwartych prekodowanych

Problematyczna z punktu widzenia rzetelności danych była też decyzja o zastosowaniu pytań, które brzmią jak pytania otwarte, a mają zamkniętą kafeterię odpowiedzi. Te tzw. pytania otwarte prekodowane dopuszcza się, gdy można przewidzieć odpowiedzi i stworzyć wyczerpującą kafeterię, czyli w odniesieniu do kwestii dających się ująć numerycznie (np. wiek) lub zestandaryzowanych (np. poziom wykształcenia) (Brill 2008). Natomiast w profilowaniu zastosowano je do badania spraw złożonych: np. przyczyn utrudniających podjęcie pracy lub czynności, które osoba jest w stanie zrobić w celu zwiększenia swoich szans na podjęcie pracy [MPiPS1]. Choć zdecydowano się rozbudować kafeterię na te dwa pytania aż do 23 i 13 odpowiedzi, to nadal nie była wyczerpująca. Nie było opcji pozwalających uwzględnić przyczyny utrudniające podjęcie pracy, spontanicznie wymieniane przez bezrobotnych podczas profilowania: tj. brak znajomości, nieatrakcyjny wygląd, bezdomność, karalność [MPiPS2: 51–52]. Z opcji „inne", zrezygnowano, prawdopodobnie dlatego, że nie można jej było poddać zautomatyzowanej obróbce.

Zastosowanie pytań prekodowanych wymusiło dużą uznaniowość urzędników. Ich zadanie polegało na przełożeniu nieustrukturyzowanej wypowiedzi profilowanego na przewidziane kategorie. Zadanie to utrudniał słabej jakości kwestionariusz, brak jednolitych wytycznych jak prowadzić profilowanie i postępować w sytuacjach nieprzewidzianych na liście odpowiedzi (61% urzędników spotkało się z taką sytuacją w co najmniej połowie przypadków profilowania) [CAWI]. W efekcie każdy profilujący wypracowywał sobie swój sposób zadawania pytań i interpretacji odpowiedzi. Zaowocowało to zróżnicowaniem praktyk: od ścisłego trzymania się kwestionariusza do zmieniania kolejności pytań (62% ankietowanych zmieniało ich kolejność w co najmniej połowie przypadków profilowania) i przeformułowywania ich (82% ankietowanych przeformułowało pytania w co najmniej połowie przypadków profilowania) [CAWI].

Jeśli odpowiedzi udzielane doradcy klienta mogły być przydatne w pracy z konkretną osobą bezrobotną, to sensowność poddania ich statystycznej obróbce była wątpliwa. Z powodu niskiej jakości kwestionariusza, przymusowego charakteru profilowania oraz różnic w przeprowadzaniu ankiety uzyskane dane nie były ani zestandaryzowane, ani porównywalne, tylko stwarzały takie złudzenie.

2. Operacje na danych

2.1 Zastosowanie modelu psychometrycznego

Do statystycznego profilowania najczęściej stosuje się modele predykcyjne (Allhutter et al. 2020; Desiere et al. 2019; Wiśniewski, Wojdyło-Preisner 2013, 2015). Profilowanie polega wtedy na obliczaniu prawdopodobieństwa, z jakim osoba przez dłuższy okres nie znajdzie pracy. W zależności od otrzymanego

wyniku przypisuje się ją do jednej z grup ryzyka. W takich statystycznych modelach predykcyjnych jakość analiz jest weryfikowalna empirycznie: w fazie testowania dokonuje się wyboru zmiennych i technik analitycznych, które najefektywniej przewidują ryzyko bezrobocia w ramach konkretnego zbioru danych. Jest to możliwe, ponieważ zmienna zależna (tj. określony czas pozostawania bez pracy) jest obserwowalna.

W Polsce – o czym świadczy zastosowanie wskaźnika Alfa Cronbacha (Kline 2005: 167) – zdecydowano się na inne narzędzie: model psychometryczny. Takie modele służą do pomiaru zmiennych ukrytych (czyli nieobserwowalnych), a więc pewnych wewnętrznych właściwości przypisywanych jednostkom. Inaczej mówiąc, za pomocą narzędzi psychometrycznych mierzy się konstrukty, których istnienie i oddziaływanie się jedynie zakłada. Konstruktów nie można jednak obserwować bezpośrednio, a jedynie za pośrednictwem innych obserwowalnych zmiennych, uznanych za ich empiryczne wskaźniki. Przykładem jest pomiar inteligencji lub zdolności matematycznych (zmienne ukryte), którego dokonuje się testując, jak dana osoba rozwiązuje zadania (empiryczne wskaźniki).

Powyższa charakterystyka modeli psychometrycznych ma istotne znaczenie dla oceny ich jakości pomiarowej. Ponieważ jedynie przyjmuje się, że mierzone konstrukty mają wpływ na obserwowalne empirycznie zjawiska (tak jak zakłada się, że inteligencja lub zdolności matematyczne mają wpływ na umiejętność rozwiązywania zadań), kluczowa staje się rzetelna konceptualizacja takich konstruktów (Babbie 2004: 148). Innymi słowy trzeba zdefiniować, jaką ukrytą właściwość jednostek chce się mierzyć i ustalić jej spójną charakterystykę. W ramach tego procesu wyeksplikowany powinien zostać zakres znaczeniowy oraz założenia dotyczące badanego, nieobserwowalnego zjawiska. Tylko w ten sposób można odpowiedzieć na pytanie, co dane narzędzie ma mierzyć i czy są przesłanki, by zakładać, że ta ukryta właściwość oddziałuje empirycznie.

Niepoprawna konceptualizacja

W przypadku zastosowanego w Polsce modelu mierzonej zmiennej ukrytej nadawano różne nazwy: potencjał aktywności zawodowej, aktywizacyjny, zatrudnieniowy [MPiPS2; 8; 10]. Wydaje się, że twórcy narzędzia nie mieli jasności, jak nazywać nieobserwowalną właściwość, która stała się podstawą nowego sposobu klasyfikacji osób bezrobotnych. Problem był jednak poważniejszy niż niespójne nazewnictwo – dotyczył poprawności konceptualizacji.

W analizowanym przypadku potencjał zatrudnieniowy (tego określenia będziemy używać dla uproszczenia) skonceptualizowano poprzez dwa wymiary, również konstrukty (nieobserwowalne własności): oddalenie od rynku pracy (O) oraz gotowość do podjęcia lub powrotu na rynek pracy (G), które badano za pośrednictwem pytań w kwestionariuszu (empiryczne wskaźniki). W dokumentach

statystycznych G definiowano jako „elementy ogólnej postawy osoby bezrobotnej wobec poszukiwania zatrudnienia", „czynniki psychologiczne, światopoglądowe i osobiste, określające poziom zmotywowania osoby bezrobotnej do podjęcia pracy" [MPiPS3: 3–4]. Natomiast O określono jako „funkcję cech determinujących szansę otrzymania (…) oferty pracy i utrzymania zatrudnienia (demografia, kwalifikacje, możliwości dojazdu do pracy itp.)" [MPiPS5: 4] lub „czynniki utrudniające bezrobotnemu wejście lub powrót na rynek pracy" [MPiPS8: 6]. Jak uzasadniano „wymiary te świadomie rozdzielono już na etapie konceptualizacji modelu", zakładając, że na potencjał zatrudnieniowy składa się zarówno wewnętrzna motywacja danej osoby (G), jak i czynniki pozamotywacyjne stanowiące barierę do podjęcia pracy (O) [MPiPS3: 3].

Kierunek zależności przyczynowo-skutkowej – sprzeczność założeń modelu z danymi

Każdy model psychometryczny zakłada, że zależność przyczynowo-skutkowa przebiega od konstruktu (przyczyna) do zmiennych wskaźnikowych (skutek). Ten kierunek zależności jest dość intuicyjny w typowych zastosowaniach psychometrii. Na przykład, jeżeli mierzy się zdolności matematyczne czy inteligencję, wówczas przyjmuje się, że to od tych konstruktów zależy, jak osoba rozwiązuje zadania.

Problematyczność konceptualizacji potencjału zatrudnieniowego[11] zaczyna być widoczna, gdy przyjrzymy się, co uznano w profilowaniu za empiryczne wskaźniki O: były nimi cechy społeczno-demograficzne (np. wiek, płeć, wykształcenie) czy inne uwarunkowania utrudniające podjęcie pracy (np. ograniczenia zdrowotne, opieka nad dziećmi, wielkość miejscowości, możliwość dojazdu). Stosując model psychometryczny milcząco założono więc, że oddalenie od rynku pracy jest predyspozycją jednostki (przyczyna), która oddziałuje na te cechy strukturalne (skutek). Tak jakby osoby bezrobotne miały ukrytą wewnętrzną właściwość O, której wartość wpływa między innymi na to, ile mają lat, jakie mają wykształcenie, gdzie mieszkają. Oddalenie, a zatem również potencjał zatrudnieniowy, były więc nieprawidłowo skonceptualizowane.

Brak testu trafności

Jakość pomiarowa każdego narzędzia zależy od trafności pomiaru. Trafność mówi nam, na ile narzędzie faktycznie mierzy zmienną, którą ma za zadanie uchwycić (por. Babbie 2004: 166). Zgodnie ze standardami metodologicznymi należało sprawdzić, czy model rzeczywiście mierzy potencjał zatrudnieniowy. Trudność polegała na tym, że ów potencjał jest nieobserwowalny – został

[11] W tym miejscu pomijamy temat założeń normatywnych na temat bezrobotnych, które leżą u podstaw takiej konceptualizacji. Więcej zob. (Sztandar-Sztanderska and Zieleńska 2020) .

powołany do życia za pośrednictwem konceptualizacji. Wiemy już, że konceptualizacja była przeprowadzona nieprawidłowo: założono błędny kierunek zależności między zmiennymi. Uniknięcie tego błędu było możliwe, gdyby przeprowadzono odpowiadający standardom test trafności narzędzia. Tak się jednak nie stało.

W psychometrii dominującym podejściem do weryfikacji trafności jest walidacja konstruktu (Furr, Bacharach 2013: 216). Oznacza to, że należy przeprowadzić testy sprawdzające, czy wyniki pomiaru są w sposób satysfakcjonujący powiązane z mierzonym konstruktem, który z kolei powinien mieć podstawy w teorii. Jednym ze sposobów walidacji potencjału zatrudnieniowego byłoby więc postawienie hipotez, (najlepiej) opierając się na teoretycznych przesłankach dotyczących spodziewanych kierunków zależności tej zmiennej z innymi zmiennymi, a następnie zweryfikowanie ich empirycznie. Z jednej strony hipotezy powinny dotyczyć tego, z jakimi zmiennymi skorelowany jest pozytywnie lub negatywnie potencjał zatrudnieniowy (tzw. trafność zbieżna). Z drugiej strony hipotezy powinny dotyczyć tego, z jakimi zmiennymi nie będzie korelował ten test (trafność różnicowa) (por. Cohen, Swerdlik 2009: 189 i dalej)[12].

W przypadku profilowania nie zastosowano jednak walidacji konstruktu. Natomiast uznano, że narzędzie spełnia swoją funkcję, skoro algorytm przypisuje bezrobotnych do profili w sposób nie budzący „masowego sprzeciwu pracowników Urzędów Pracy" [MPiPS6: 3]. Proponowano by „polegać na [ich] intuicji" i traktować korekty profilu jako sygnał, że narzędzie dokonało niewłaściwego pomiaru i kategoryzacji [MPiPS3: 26], a takich korekt na etapie kalibrowania algorytmu było niewiele: zaledwie 0,58% wszystkich przypadków profilowań [MPiPS4: 43][13]. Innymi słowy, zgodę pracowników PUP na automatycznie generowane wyniki potraktowano jako odpowiednik testu trafności. Z dokumentów statystycznych wynika jednak, że zdawano sobie sprawę, że pracownicy PUP unikają wprowadzania korekt [MPiPS4: 7], co podważało tę – i tak wątpliwą metodologicznie – argumentację[14].

Prawidłowo wykonany test trafności narzędzia pomiarowego, a więc choćby pośrednia weryfikacja, czy zmienne obserwowalne mierzą dany konstrukt,

[12] Jest to jeden z wielu przykładów walidacji konstruktu.

[13] Rozważano też, czy nie wprowadzić zmian do algorytmu, inspirowanych dokonywanymi wcześniej przez urzędników korektami. Zastanawiano się, czy przeprogramować algorytm, tak by próbował wystąpienie takich korekt przewidzieć i dokonywał w tych sytuacjach innego zaszeregowania [MPiPS4]. Z raportów statystycznych nie wynika jednak, by takie rozwiązanie wprowadzono. Natomiast wcześniejsze korekty brano pod uwagę jako jeden z powodów, uzasadniających zmianę sposobu wyznaczania granic między profilami [MPiPS5].

[14] Zdając sobie z tego sprawę MPiPS podjęło działania, które miały zachęcać urzędników do wprowadzania korekt, gdy uznają wynik za nietrafny, m.in. dodano listę z zestandaryzowanych uzasadnień decyzji o zmianie profilu [MPiPS11].

jest najważniejsza w sytuacji, gdy na wynikach testów oparte są istotne decyzje (Furr, Bacharach 2013: 203). Jest to rudymentarna wiedza, przekazywana w podręcznikach: „bez walidacji decyzje oparte na testach dotyczące osób mogą być oparte na błędnych informacjach, a nawet szkodliwe. (…) Takie decyzje mogą potencjalnie wpłynąć na życie osób nimi dotkniętych, a trafność testu może mieć istotne znaczenie dla tych decyzji" (Furr, Bacharach 2013: 203). Dlatego tak niepokojące jest zlekceważenie tej kwestii. W efekcie tak naprawdę nie wiemy, co mierzono podczas profilowania. Wiemy natomiast, że wyimaginowany, nieprawidłowo skonceptualizowany konstrukt stał się podstawą klasyfikacji bezrobotnych, rzutującej na dystrybucję ALMP.

2.2 Błędne zastosowanie Alfy Cronbacha jako miary rzetelności

Drugą problematyczną decyzją był wybór Alfy Cronbacha do oceny rzetelności. Alfa Cronbacha mierzy wewnętrzną spójność skal. Przyjmuje się, że stosunkowo wysoka wartość wskaźnika (powyżej 0,7) oznacza, że pytania w ramach danej skali mierzą to samo zjawisko (Kline 2005:182; Lavrakas 2008). W przypadku narzędzia profilowania zastosowano ten wskaźnik, by ocenić wewnętrzną spójność skal G i O, choć jako satysfakcjonujący uznawano wynik już powyżej 0,6 [MPiPS: 3]. Konstrukty G i O były mierzone za pomocą listy pytań, do których przypisano różnie punktowane odpowiedzi. Przyjęto, że im więcej punktów uzyskała osoba bezrobotna, tym większe jej oddalenie od rynku pracy (O) i tym mniejsza gotowość do podjęcia pracy (G). Za pomocą Alfy Cronbacha sprawdzano, czy zestawy odpowiedzi w ramach każdej ze skal są ze sobą pozytywnie skorelowane: odpowiednio wysoka wartość tego współczynnika miała świadczyć o rzetelności pomiaru.

Jednak by prawidłowo zastosować Alfę Cronbacha musi być spełnionych szereg założeń wynikających zarówno z klasycznej teorii testu (KTT) (Kline 2005:167–168), jak i założeń Alfy Cronbacha jako miary rzetelności (McNeish 2018; Sijtsma 2009). Założenia związane z Alfą uznawane są w literaturze przedmiotu za trudne do spełnienia czy wręcz nierealistyczne, stąd rekomenduje się stosowanie innych wskaźników rzetelności niż Alfa (McNeish 2018; Steinberg, Thissen 1996). W analizowanym przez nas przypadku tych założeń nie spełniono, co wyjaśniamy poniżej. W efekcie Alfa Cronbacha szacowała rzetelność narzędzia nieprecyzyjnie. Jest to znaczące, gdyż naukową legitymizację narzędzia opierano właśnie na satysfakcjonującej wartości Alfy Cronbacha (sic!).

KTT stworzona została na potrzeby badania nieobserwowalnych konstruktów. Jej najbardziej podstawowe założenie mówi więc, że każdy pomiar zmiennej ukrytej obarczony jest błędem, gdyż nie mamy możliwości obserwować bezpośrednio jej prawdziwej wartości. Wyobraźmy sobie to na przykładzie: uczniowie badani pod kątem zdolności matematycznych otrzymują zestaw zadań do rozwiązania. Wynik uzyskany przez ucznia w każdym z zadań

traktowany jest jako jeden ze wskaźników owych zdolności. Zgodnie z KTT przyjmuje się, że otrzymany wynik, tzw. wartość obserwowana (Z) zależy z jednej strony od faktycznych zdolności ucznia (T, tzw. *true score*), z drugiej od błędu pomiaru (E, tzw. *error term*), który może być związany z nastrojem, zmęczeniem lub innymi przeszkodami o charakterze losowym. Model taki zapisać można formalnie jako: Z=T+E. Zakłada się także, że błędy pomiaru mają rozkłady normalne o średniej zero. Oznacza to, że osoba może z takim samym prawdopodobieństwem uzyskać wynik przeszacowujący lub niedoszacowujący jej faktyczne zdolności.

W przypadku narzędzia profilowania problem polegał jednak na tym, że przy podziale bezrobotnych na profile zignorowano błąd pomiaru wpisany w miary G i O. Innymi słowy przyjęto, że otrzymany przez osobę bezrobotną wynik jest wartością prawdziwą, a nie wartością obserwowaną, która zawiera w sobie błąd pomiaru. W rezultacie przypisanie osoby do profilu mogło być dalece nieprecyzyjne (np. zamiast do profilu I klasyfikowano kogoś do profilu II itd.). Jest to tym bardziej znaczące, że klasyfikacja do danego profilu otwierała lub zamykała dostęp do konkretnych programów rynku pracy.

Z kwestią błędu pomiaru wiąże się jednak jeszcze bardziej fundamentalny problem. Gdy przyłożymy to założenie do wskaźników zmiennej O, okaże się ono bezsensowne. Oznaczałoby to, że pomiar wskaźników takich jak wiek bezrobotnego, poziom wykształcenia czy miejsce zamieszkania jest obarczony losowym błędem pomiaru. Podejście, które z powodzeniem stosowane jest w testach psychologicznych czy pomiarach zdolności matematycznych, przyłożono do zmiennych zdających sprawę z faktów. Te zmienne zazwyczaj nie są obarczone błędem pomiaru, ponieważ można stosunkowo łatwo stwierdzić, ile dana osoba ma lat, jakie ma wykształcenie, gdzie mieszka. Tym samym nie możemy też w tym przypadku mówić o normalnym rozkładzie błędu pomiaru. Oznacza to, że zastosowano podejście analityczne, które nie odpowiada charakterowi analizowanych zjawisk.

Kolejnym założeniem Alfy Cronbacha, którego nie spełniono, jest jednowymiarowość konstruktu, czyli nieskorelowanie błędów pomiaru (por. McNeish 2018; Steinberg, Thissen 1996). Odnosząc się do przykładu testu zdolności matematycznych, jednowymiarowość oznaczałaby, że wartości punktowe uzyskane przez uczniów za poszczególne zadania zależą wyłącznie od ich zdolności matematycznych (Z). Zmienna Z byłaby jedyną determinantą tych wartości. Mówiąc językiem statystycznym, Z wyjaśniałoby całą strukturę zależności między tymi zmiennymi, a to co nie jest powiązane z Z, byłoby losowym błędem. Błędy nie mogłyby być więc skorelowane.

Zastosowane w profilowaniu wskaźniki oddalenia (O) wydają się powiązane i nie ma podstaw, by zakładać, że związek ten jest efektem O. Przykładowo, można przypuszczać, że występuje związek między następującymi wskaźnikami

O: miejsca zamieszkania z możliwością dojazdu do pracy, płci z ograniczoną dyspozycyjnością ze względu na nierówno rozłożone obowiązki opiekuńcze, wykształcenia ze znajomością języków obcych itd. Założenie o jednowymiarowości konstruktu O i nieskorelowaniu błędów pomiaru nie zostało w tej sytuacji spełnione. Jeśli zaś konstrukt jest wielowymiarowy, to wskaźnik Alfy Cronbacha nieprecyzyjnie szacuje rzetelność: albo ją zaniża, albo zawyża (nawet o 20%) (Gessaroli, Folske 2002). Oznacza to, że pomiar wartości O u osób bezrobotnych może być mniej rzetelny niż wynikałoby to z obliczonej przez twórców narzędzia wartości Alfy Cronbacha (Bentler 2009; McNeish 2018).

Kolejnym niespełnionym założeniem Alfy Cronbacha jest tzw. *tau-equivalence* (McNeish 2018). Oznacza ono, że dla osoby badanej wszystkie wskaźniki danego konstruktu mają identyczną wartość prawdziwą (T). Na przykład, osoba której wartość zmiennej zdolności matematyczne (Z) wynosiłaby cztery (T=4), we wszystkich pytaniach wskaźnikowych związanych z Z, powinna uzyskiwać wartość punktową cztery, obarczoną losowym błędem (E) zawyżającym lub zaniżającym tę wartość. Tym samym jedyne różnice w punktach przy wskaźnikach Z powinny wynikać z błędu pomiaru.

W kwestionariuszu profilowania tak ustalono punktację odpowiedzi, że osoba poddana profilowaniu nie mogła nawet uzyskać tyle samo punktów za odpowiedzi dotyczące O czy G. Np. na skali O przewidziano od 0 do 1 punktu za odpowiedzi w pytaniu o płeć, od 0 do 5 punktów w pytaniu o wiek, od 0 do 8 punktów w pytaniu o wykształcenie [MPiPS1]. Innymi słowy wartości punktowe różnicował nie tylko błąd pomiaru. Niespełnienie założenia o *tau equivalence* również prowadzi do nieprecyzyjnego oszacowania rzetelności za pomocą Alfy Cronbacha (Revelle 2020; Sijtsma 2009). W tej sytuacji Alfa Cronbacha nie doszacowuje rzetelności pomiaru.

Ostatnie niespełnione założenie Alfy Cronbacha dotyczy rodzajów skal pomiarowych, na których określono zmienne obserwowalne. Alfę Cronbacha oblicza się na podstawie korelacji Pearsona między zmiennymi obserwowalnymi manifestującymi dany konstrukt. By dokonać tego obliczenia prawidłowo, zmienne powinny być określone co najmniej na skali interwałowej. Problem polega jednak na tym, że większość zmiennych obserwowalnych opartych było na skalach porządkowych (np. stopień niepełnosprawności) lub nominalnych (np. powody rejestracji w PUP). Korelacje wyliczono zaś na podstawie wartości punktowych, arbitralnie przypisanych poszczególnym odpowiedziom (więcej na temat arbitralności punktacji dalej). Innymi słowy, analizowane zmienne nie były określone na skalach interwałowych, co najwyżej to symulowały. Tymczasem, Alfa Cronbacha obliczona w taki sposób może być niedoszacowana (McNeish 2018; Sijtsma 2009).

Podsumujmy: wybór wskaźnika Alfy Cronbacha do oceny rzetelności pomiaru był nieprawidłowy, ponieważ zignorowano założenia KTT i Alfy Cronbacha.

W konsekwencji, wartość tego wskaźnika, na którą powoływało się M(R)PiPS, legitymizując narzędzie profilowania, była nieprecyzyjnie obliczona. Nie wiadomo więc, czy – jak twierdziło M(R)PiPS – osiągnęła poziom uznawany za akceptowalny w psychometrii.

2.3 Maksymalizacja Alfy Cronbacha, czyli pierwsza odsłona torturowania danych

W trakcie analizy algorytmu profilowania nasze wątpliwości wzbudziła również nieintuicyjna wartość punktów przypisanych poszczególnym odpowiedziom kwestionariusza. Na przykład: bezrobotny otrzymywał aż 7 punktów za brak znajomości języków obcych, co miało świadczyć o jego dużym oddaleniu od rynku pracy (w najwyżej punktowanych pytaniach można było dostać maksymalnie 8 pkt). [MPiPS1]. Dla porównania: niższe wartości O przypisano cechom, o których wiadomo, że wpływają negatywnie na pozycję na rynku pracy (np. Wiśniewski, Wojdyło-Preisner 2013: 153) – takim jak długotrwałe bezrobocie (2 punkty za pozostawanie bez pracy od 12 do 24 miesięcy, a 6 punktów co najmniej 24 miesięcy) lub wiek (1 punkt za wiek poniżej 25 roku życia, 5 punktów za wiek powyżej 50 roku życia). Sprawdziliśmy więc, jak ustalano punktację odpowiedzi. Interesowało nas, czy mniejsza lub większa liczba punktów, które w ramach obu skal sumowano, miała związek z odpowiednio mniejszym lub większym oddziaływaniem danego czynnika na status danej osoby na rynku pracy.

Okazało się, że punktację przekształcano tak, by sztucznie zwiększać wartości Alfy Cronbacha wyliczane dla skal O i G – czego dowodzi obszerna część raportów statystycznych [MPiPS3: 18–24, MPiPS5: 10–23, MPiPS6: 10–21, MPiPS7: 9–21]. Dlaczego tak robiono? Zmiany punktacji miały zwiększyć wartość Alfy Cronbacha. To z kolei było potrzebne, by uznać narzędzie za rzetelne. W psychometrii przyjmuje się, że wskaźnik ten powinien wynosić powyżej 0,7 (Kline 2005: 182; Lavrakas 2008). Twórcy narzędzia za satysfakcjonujący wynik uznali wartość powyżej 0,6 [MPiPS3].

Problem polega na tym, że jednocześnie abstrahowano od faktycznej siły różnych czynników, wpływających na szanse na podjęcie pracy. W efekcie mniejsza lub większa wartość zmiennej O otrzymana za dane pytanie nie musiała odzwierciedlać większych lub mniejszych szans danej osoby na znalezienie pracy. Podobnie było ze zmienną G. Podsumowując, ustalanie punktacji służyło wpasowaniu danych w model, co w literaturze metodologicznej nazywane jest torturowaniem danych: „dane dowiodą czegokolwiek, co chciałby dowieść badacz, jeśli będzie się nimi manipulować na wystarczająco wiele sposobów" (Mills 1993:196). W tym przypadku dane miały dowieść, że skale O i G są rzetelne, i w końcu po licznych modyfikacjach zwiększono rzetelność G, ale nie do

końca udało się to zrobić w przypadku O[15] [MPiPS5]. Zastosowana strategia nie opierała się na jednym z podejść zalecanych w metodologii: czy to na podejściu konfirmacyjnym (testującym, czy obserwowane dane empiryczne mogły być wygenerowane przez założony model), czy to na podejściu eksploracyjnym (testującym dopasowanie różnych modeli do danych, by wybrać najlepszy z nich). Od początku brano pod uwagę jedynie model Alfa Cronbacha i by móc uznać wynik wskaźnika za rzetelny manipulowano punktacją.

2.4 Podział na trzy profile, czyli druga odsłona torturowania danych

Nasze zastrzeżenia budzi też sposób podziału osób bezrobotnych pomiędzy profile. Przed rozpoczęciem analiz założono wynik, do którego ma doprowadzić opracowywane narzędzie i do niego dopasowywano sposób wyznaczenia granic między profilami. Osoby bezrobotne miały być podzielone na trzy, hierarchicznie uporządkowane grupy (ze względu na wartości zmiennych G i O) i – co niezwykle istotne – bez żadnych obliczeń przyjęto w jakich proporcjach ten podział powinien być dokonany przez algorytm (w jednym wariancie miało to być: I profil – 20%; II – 60%; III – 20%; w drugim: 15% – 70% – 15%) [MPiPS3: 26, MPiPS6: 8, PM1].

Te wstępne założenia wyznaczały ograniczone ramy dla analiz statystycznych. Najbardziej problematyczne i skutkujące torturowaniem danych było wstępne założenie o proporcjach między profilami. W standardowo stosowanych technikach klasyfikacji (np. hierarchicznym klastrowaniu), podziału dokonuje się grupując jednostki podobne do siebie: proporcje tych grup są wtedy wynikiem analizy empirycznej. W przypadku narzędzia profilowania postąpiono odwrotnie. Nie zakreślano granic między profilami w sposób, który pozwoliłby wyłonić – jak deklarowano w oficjalnych dokumentach – „homogeniczne grupy", zbliżone pod kątem problemów czy potrzeb [MPiPS8: 4]. Metod statystycznych nie użyto do diagnozy, ile osób znajduje się w trudnej, ile w umiarkowanej, a ile w dobrej sytuacji. Natomiast wykorzystując dane pochodzące z pilotażu (potem skorygowane o dane z pięciu miesięcy profilowania [MPiPS5]), ustalono formułę matematyczną, która pozwoliłaby wyznaczyć granice między profilami tak by osiągnąć „właściwą" [MPiPS5] (czytaj: założoną odgórnie) proporcję. Gdyby tylko dane o bezrobotnych wprowadzane w kolejnych miesiącach i latach przez pracowników

[15] Z pozyskanych dokumentów wynika, że od pilotażu, zrealizowanego w 2013 r. wartość Alfy Cronbacha dla skali G zwiększono z 0,395 do powyżej 0,7 i ta wartość utrzymywała się zarówno w maju, sierpniu i grudniu 2014 r. [MPiPS3, MPiPS5, MPiPS6, MPiPS7, MPiPS8] W przypadku skali O aż do grudnia 2014 r. nie udało się osiągnąć nawet wartości 0,6. Natomiast, na podstawie wykonanej symulacji zakładano, że – przy zmianie punktacji odpowiedzi – uda się osiągnąć wynik Alfa Cronbacha na poziomie 0,661 [MPiPS5]. Większość tych zmian wprowadzono [MPiPS1], ale nie wiadomo, jaką faktyczną wartość osiągał później ten wskaźnik, ponieważ M(R)PiPS nie udostępniało na ten temat późniejszych informacji.

PUP były zbliżone do danych wejściowych, na podstawie których opracowano algorytm, podział między profilami wynosiłby stale: 15%–70%–15%[16].

Podsumowanie

W artykule podjęliśmy problem jakości poznawczej algorytmów stosowanych w polityce społecznej na przykładzie narzędzia profilowania bezrobotnych wykorzystywanego w Polsce w latach 2014–2019. Interesowało nas, jakie były podstawy przypisywanych przez algorytm klasyfikacji. Czy dane, na których opierał się pomiar, były wiarygodne? Co było przedmiotem pomiaru? Czy pomiar był trafny i rzetelny? Skąd brały się błędy? W wyniku przeprowadzonej analizy odkryliśmy, że algorytm profilowania nie spełniał nawet najbardziej podstawowych standardów metodologicznych, zarówno jeśli chodzi o rzetelność danych, jak i dokonywane nań operacje statystyczne.

Po pierwsze, okazało się, że kwestionariusz stosowany do zbierania danych o bezrobotnych zawierał podręcznikowe błędy: część pytań była niejednoznaczna, niezrozumiała lub sugerująca odpowiedzi (Babbie 2004), a pracownicy PUP w sposób zróżnicowany prowadzili wywiad kwestionariuszowy. W efekcie dane o bezrobotnych przetwarzane przez model były nieporównywalne, niewystandaryzowane, co podaje w wątpliwość ich rzetelność. O ile odpowiedzi cząstkowe mogły być przydatne dla doradcy klienta PUP w pracy z konkretną osobą bezrobotną, o tyle sensowność poddania ich statystycznej obróbce była wątpliwa. Zgodnie z ustaleniami metodologii badań ilościowych i badań o algorytmach, zakładamy, że błędy w danych przełożyły się na błędne wyniki i słabej jakości algorytm (Mittelstadt et al. 2016).

Po drugie, zastosowany model nie miał charakteru predykcyjnego, lecz dokonywał pomiaru zmiennej ukrytej, której nie skonceptualizowano w sposób przyjęty w badaniach społecznych czy psychometrii. Wymyślono, że wszystkich bezrobotnych cechuje pewna ukryta właściwość, której nadawano różne nazwy: potencjał aktywności zawodowej, aktywizacyjny czy zatrudnieniowy. Nie zdefiniowano jednak poprawnie, czym ta właściwość jest i nie upewniono się, czy pomiar jest trafny. W praktyce oznacza to, że nie wiadomo, co mierzył algorytm profilowania używany we wszystkich urzędach pracy. Skoro zaś nie zdefiniowano, co jest przedmiotem pomiaru, to nie można było też zweryfikować precyzji pomiaru, czy korygować skrzywień krzywdzących dla określonych grup społecznych (*bias*).

[16] Tak się jednak nie stało i proporcje kształtowały się inaczej: w latach 2015–2019 osoby zaklasyfikowane do I profilu stanowiły około 2%, osoby w II profilu między 63 a 68%, a osoby w III profilu między 29 a 35% (dane według stanu na 31 grudnia każdego roku) [MRPiPS2].

Po trzecie, zastosowano model psychometryczny oparty na klasycznej teorii testu, ignorując podstawowe założenia, na których się ten model opiera. Pomylono też kierunek zależności przyczynowo-skutkowych. Zgodnie z logiką modelu przyjęto, że od wewnętrznej właściwości osób bezrobotnych będzie zależeć – by przypomnieć najbardziej absurdalne przykłady – wykształcenie danej osoby lub jej miejsce zamieszkania. Takie założenia choć nie zostały wypowiedziane, stanowiły fundament wdrożonego systemu.

Po czwarte, z naszej analizy wynika, że ostatecznie przyjęta formuła matematyczna i wartości zmiennych nie były efektem analiz empirycznych, lecz dopasowywano je tak, by generowane przezeń wyniki spełniły z góry przyjęte założenia. To zjawisko nazywa się w literaturze „torturowaniem danych" (Mills 1993). W praktyce oznaczało to, że opracowując algorytm skupiono się na tym, by generował on z góry założone proporcje między osobami z profilu I, II i III, a zignorowano, czy w ten sposób tworzone grupy są homogeniczne pod kątem sytuacji na rynku pracy lub potrzeb. Ponadto manipulując punktacją, starano się zmaksymalizować Alfę Cronbacha – wskaźnik rzetelności skal używany do legitymizacji narzędzia. W efekcie wartości zmiennych nie musiały odzwierciedlać szans danej osoby na znalezienie pracy.

Przedstawione w artykule wyniki dostarczają mocnych argumentów na rzecz otwierania czarnej skrzynki technologii i weryfikowania jakości modeli matematycznych zintegrowanych z oprogramowaniem komputerowym przed ich wykorzystaniem w polityce społecznej (Citron, Pasquale 2014; O'Neil 2017). Choć ten postulat nie jest nowy, to zestawiając typy błędów poznawczych znane z literatury przedmiotu z wynikami naszych badań, twierdzimy, że prawdopodobnie nie docenia się, jak niskiej jakości mogą być algorytmy stosowane w polityce społecznej oraz jak – z perspektywy metodologii – podstawowe błędy popełnia się przy ich konstrukcji i przy zbieraniu danych. Przypomnijmy, że w analizowanym przez nas przypadku problemem było coś tak fundamentalnego jak brak konceptualizacji zmiennej ukrytej, która stała się podstawą klasyfikacji osób bezrobotnych.

Na koniec chcieliśmy zwrócić uwagę na dwie bariery, utrudniające zdobywanie wiedzy o algorytmach stosowanych w polityce społecznej oraz zaproponować możliwe rozwiązania. Pierwszą barierę stanowi odmowa dostępu do algorytmu, którą uzasadnia się przeciwdziałaniem nadużyciom (Pasquale 2015: 4). Z tego względu niezwykle istotne wydaje nam się budowanie strategii badawczej, opartej na współpracy z sygnalistami pracującymi w instytucjach publicznych oraz z aktywistami z organizacji pozarządowych. *Case* profilowania pokazuje, że dostęp do algorytmu można uzyskać na drodze sądowej, powołując się na prawo obywateli do informacji publicznej – tak jak uczynili to prawnicy z Fundacji Panoptykon, używając w swojej argumentacji materiałów, które wyciekły wcześniej z urzędów pracy. Inna strategia – którą udało nam się tylko

częściowo wdrożyć dzięki współpracy ze strony MRPiPS – polega na budowaniu u decydentów politycznych i urzędników przekonania o potrzebie niezależnych i długofalowych badań algorytmów opartych na udostępnionych przez nich danych. Ważne jest, by pracownicy instytucji publicznych z dystansem podchodzili do zapewnień o skuteczności administracji, która wykorzystuje algorytmy oraz trafności i obiektywności algorytmicznych diagnoz. Ta retoryka jest niebezpieczna, ponieważ legitymizuje zastosowanie narzędzi, których jakość – jak dowodzi przypadek profilowania – może być mierna.

Drugą przeszkodę na drodze lepszego poznania algorytmów stosowanych w polityce publicznej jest niedoreprezentacja badań prowadzonych poza Stanami Zjednoczonymi i krajami wysoko rozwiniętymi. O ile w literaturze przedmiotu dużo się pisze o wyrafinowanych narzędziach, stosujących uczenie maszynowe (*machine learning*) i przeszukujących wielkie zbiory danych, o tyle wciąż niewiele jest analiz algorytmów wdrażanych w krajach, które trudno uznać za technologiczne centrum. Trzeba dopuszczać możliwość, iż opisy zagrożeń i błędów, które znajdujemy w literaturze, to jedynie czubek góry lodowej. Całą górę moglibyśmy lepiej poznać, uwzględniając w badaniach kraje (pół)peryferyjne, niedofinansowane sektory polityki społecznej, rozwiązania kierowane do stygmatyzowanych grup docelowych, które nie dysponują możliwością wywierania nacisku politycznego (Eubanks 2017; O'Neil 2017). W tego rodzaju badaniach warto zrekonstruować proces podejmowania decyzji oraz rozpoznać instytucjonalne, organizacyjne czy jednostkowe przyczyny błędów.

Spis analizowanych dokumentów

[GAZETA_PRAWNA] Karolina Topolska, Męcina o Efektach Reformy: Urzędy Pracy Muszą Być Efektywne, *Gazeta Prawna*, 23.09.2014

[INTERPELACJA] Ewa Kołodziej (2016) Interpelacja poselska nr 2402 do MRPiPS w sprawie oceny funkcjonowania w praktyce instytucji zapewnionych przez Ustawę o promocji zatrudnienia i instytucjach rynku pracy w kontekście osób bezrobotnych i trwale

[MPiPS1] MPiPS (2014) *Kwestionariusz profilowania z punktacją*

[MPiPS2] MPiPS (2013) *Raport z Realizacji Testu Kwestionariusza Do Profilowania Pomocy Dla Osób Bezrobotnych.*

[MPiPS3] MPiPS (2013) *Profilowanie Pomocy Dla Osób Bezrobotnych Raport z Analizy Danych z Pilotażu*, Warszawa.

[MPiPS4] MPiPS (2014) *Powody Zmiany Profilu Pomocy. Raport z Analizy Danych Jakościowych z Profilowania Osób Bezrobotnych w 2014 Roku*.

[MPiPS5] MPiPS (2014) *Profilowanie Pomocy Dla Osób Bezrobotnych Raport Końcowy z Wyników Profilowania Pomocy Dla Osób Bezrobotnych w 2014 Roku.*

[MPiPS6] MPiPS (2014) *Profilowanie Pomocy Dla Osób Bezrobotnych Raport nt. wy-*

ników Profilowania Pomocy Dla Osób Bezrobotnych Analiza Danych Cząstkowych z 319 Urzędów Pracy.

[MPiPS7] MPiPS (2014) Profilowanie Pomocy Dla Osób Bezrobotnych Raport z Analizy Danych z Profilowania w 5 Urzędach Pracy: W Warszawie, Gdańsku, Jarosławiu, Nysie i Oświęcimiu.

[MPiPS8] MPiPS (2014). *Profilowanie Pomocy Dla Osób Bezrobotnych. Podręcznik Dla Pracowników Powiatowych Urzędów Pracy*.

[MPiPS9] MPiPS (2013) *Uzasadnienie projektu ustawy o zmianie ustawy o promocji zatrudnienia i instytucjach rynku pracy oraz niektórych innych ustaw*, Druk nr 1949.

[MPiPS10] MPiPS (2012) List zapraszający wybranych ekspertów do współpracy przy projektowaniu narzędzia.

[MPiPS11] Pismo Ministra Pracy i Polityki Społecznej do Dyrektorów Powiatowych Urzędów Pracy, 3.11.2014, DRP-III-0212-149-SS/14.

[MRPiPS1] MRPiPS (2016) Odpowiedź na Interpelację poselską nr 2402 do MRPiPS w sprawie oceny funkcjonowania w praktyce instytucji zapewnionych przez Ustawę o promocji zatrudnienia i instytucjach rynku pracy w kontekście osób bezrobotnych i trwale bezrobotnych.

[MRPiPS2] MRPiPS (2020) Informacja o bezrobotnych i poszukujących pracy w grudniu 2019, Warszawa, Styczeń 2020, https://psz.praca.gov.pl/-/11341791--statystyki-strukturalne-grudzien-2019 (dostęp: 20.06.2020)

[NIK1] NIK (2019) Aktywizacja bezrobotnych – ryba czy wędka? https://www.nik.gov.pl/aktualnosci/aktywizacja-bezrobotnych-ryba-czy-wedka.html (dostęp 1.06.2020)

[NIK2] NIK (2019) Efektywność świadczenia usług rynku pracy. Informacja o wynikach kontroli, Najwyższa Izba Kontroli, Delegatura w Lublinie

[ROZPORZĄDZENIE] Minister Pracy i Polityki Społecznej (2014), Rozporządzenie Ministra Pracy i Polityki Społecznej z dnia 14 maja 2014 r. w sprawie profilowania pomocy dla bezrobotnego, Dz.U. 2014 poz. 631.

[RPO] RPO (2016) Wniosek RPO do Trybunału Konstytucyjnego w sprawie przepisów o profilowaniu danych osób bezrobotnych

[TK], TK (2018) Orzeczenie w sprawie wniosku Rzecznika Praw Obywatelskich dotyczącego zarządzania pomocą kierowaną do osób bezrobotnych.

[USTAWA] Sejm. 2014. *Ustawa z Dnia 14 Marca 2014 r. o Zmianie Ustawy o Promocji Zatrudnienia i Instytucjach Rynku Pracy Oraz Niektórych Innych Ustaw*, Dz.U. 2014 poz. 598.

[WSA] Wyrok Wojewódzkiego Sądu Administracyjnego w Warszawie z dnia 5 kwietnia 2016 r., II SAB/Wa 1012/15.

Bibliografia

Allhutter, Doris, Florian Cech, Fabian Fischer, Gabriel Grill, Astrid Mager. 2020. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data* 3: 5. DOI:10.3389/fdata.2020.00005.

Alston, Philip. 2019. Report of the Special Rapporteur on extreme poverty and human rights. United Nations Report Assembly. https://undocs.org/A/74/493.

Angwin, Julia, Jeff Larson, Surya Mattu, Lauren Kirchner. 2016. Machine Bias. Text/html. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Dostęp 6.03.2019.

Babbie, Earl R. 2003. *Badania społeczne w praktyce*. Przekład Witold Betkiewicz, Marta Bucholc, Przemysław Gadomski, Jacek Haman. Warszawa: Wydawnictwo Naukowe PWN.

Bentler, Peter M. 2009. Alpha, Dimension-Free, and Model-Based Internal Consistency Reliability. *Psychometrika*, 74, 1: 137–43. DOI:10.1007/s11336-008-9100-1.

Berman, Emily. 2018. A Government of Laws and Not of Machines. *BOSTON UNIVERSITY LAW REVIEW* 98: 1277–1355.

Bovens, Mark, Stavros Zouridis. 2002. From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control. *Public Administration Review*, 62, 2: 174–84. DOI:10.1111/0033-3352.00168.

Bowker, Geoffrey C., Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press.

Brill, Jonathan. 2008. Precoded Question. W: P. Lavrakas, ed. *Encyclopedia of Survey Research Methods*. Thousand Oaks: Sage. DOI:10.4135/9781412963947.

Citron, Danielle, Frank A. Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89 (2014–8).

Cohen, Ronald Jay, Mark Swerdlik. 2009. *Psychological testing and assessment: an introduction to tests and measurement*. 7th ed. Boston: McGraw-Hill Higher Education.

Corbanese, Valli, Gianni Rosas. 2017. Profiling youth labour market disadvantage: A review of approaches in Europe. International Labour Organization.

Criado-Perez, Caroline. 2019. *Invisible women: data bias in a world designed for men*. New York: Abrams Press.

Denzin, Norman. 1970. *The Research Act: A Theoretical Introduction to Sociological Methods*. Chicago: Aldine Pub. Co.

Desiere, Sam, Kristine Langenbucher, Ludo Struyven. 2019. Statistical Profiling in Public Employment Services: An International Comparison. OECD Social, Employment and Migration Working Papers 224. T. 224. OECD Social, Employment and Migration Working Papers. doi:10.1787/b5e5f16e-en.

Dubois, Vincent, Morgane Paris, Pierre-Edouard Weill. 2018. Targeting by Numbers. The Uses of Statistics for Monitoring French Welfare Benefit Recipients. W: L. Barrault-Stella, P.-E. Weill, eds. *Creating Target Publics for Welfare Policies*, 17: 93–109. Cham: Springer International Publishing. DOI:10.1007/978-3-319-89596-3_5.

Elish, Madeleine Clare. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5: 1–29. DOI:10.17351/ests2019.260.

Eubanks, Virginia. 2017. *Automating inequality: how high-tech tools profile, police, and punish the poor*. First Edition. New York, NY: St. Martin's Press.

Flaszyńska, Ewa. 2020. Profilowanie pomocy dla osoby bezrobotnej – nieudany eksperyment czy stracona szansa? *Praca Socjalna*, 3, 35: 109–129.

Furr, R. Michael, Verne R. Bacharach. 2013. *Psychometrics: An Introduction*. 2 edition. Lod Angeles: SAGE Publications, Inc.

Gessaroli, Marc E., Jane C. Folske. 2002. Generalizing the Reliability of Tests Comprised of Testlets. *International Journal of Testing*, 2, 3–4: 277–95. DOI:10.1080/15305058.2002.9669496.

Godlewska-Bujok, Barbara. 2020. Problem „profilowania bezrobotnych" w orzecznictwie sądów administracyjnych. *Monitor Prawa Pracy*. 1: 21-25. DOI: 10.32027/MOPR.20.1.3.

Harcourt, Bernard. 2005. Against Prediction: Sentencing, Policing, and Punishing in an Actuarial Age. *PUBLIC LAW AND LEGAL THEORY WORKING PAPER* 94.

Herman-Pawłowska, Katarzyna, Piotr Stronkowski, Stanisław Bienias, Magdalena Dybaś, Maciej Kolczyński, Justyna Kulawik-Dutkowska, Paulina Skórska. 2016. Ocena skutków regulacji wybranych aspektów wdrażania Ustawy z dnia 14 marca 2014 r. o zmianie ustawy o promocji zatrudnienia i instytucjach rynku pracy oraz niektórych innych ustaw. http://ideaorg.eu/file.php?i=podstrony/50a4f51c7d20eb-54f2e932940448c6fc.pdf.

Kline, Theresa. 2005. *Psychological testing: a practical approach to design and evaluation*. Thousand Oaks, Calif: Sage Publications.

Konarski, Roman, Michał Kotnarowski. 2007. Zastosowanie metody propensity score matching w ewaluacji ex-post. W: A. Haber, red. *Ewaluacja ex-post: teoria i praktyka badawcza*. Warszawa: Polska Agencja Rozwoju Przedsiębiorczości, 183–209.

Krosnick, Jon A., Stanley Presser. 2010. W: P. V. Marsden, J. D. Wright, eds. *Handbook of survey research*, Second edition. Bingley, UK: Emerald, 263–315.

Lavrakas, Paul. 2008. *Encyclopedia of Survey Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc. DOI:10.4135/9781412963947.

Lyon, David. 2005. *Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination.* Florence: Taylor and Francis. http://public.eblib.com/choice/publicfullrecord.aspx?p=240591.

Marshall, Thomas. 1950. *Citizenship and Social Class and other essays*. Cambridge: Cambridge University Press.

Mazur, Joanna. 2018. Right to Access Information as a Collective-Based Approach to the GDPR's Right to Explanation in European Law. *Erasmus Law Review*, 11, 3: 178–89. DOI:10.5553/ELR.000116.

McNeish, Daniel. 2018. Thanks Coefficient Alpha, We'll Take It from Here. *Psychological Methods*, 23, 3: 412–33. DOI:10.1037/met0000144.

Mileszczyk, Natalia, Bartek Paszcza, Alek Tarkowski. 2019. AlgoPolska. Zautomatyzowane podejmowanie decyzji w służbie społeczeństwu. Kraków, Warszawa: Klub Jagieloński, Centrum Cyfrowe.

Mills, James L. 1993. Data Torturing. *New England Journal of Medicine*, 329, 16: 1196–99. DOI:10.1056/NEJM199310143291613.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, Luciano Floridi. 2016. The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3, 2: 205395171667967. DOI:10.1177/2053951716679679.

Mozzana, Carlotta. 2019. A Matter of Definitions: The Profiling of People in Italian Active Labour Market Policies. *Historical Social Research / Historische Sozialforschung Vol. 44* No. 2. GESIS – Leibniz-Institut für Sozialwissenschaften: 225–246. DOI:10.12759/HSR.44.2019.2.225-246.

Niklas, Jędrzej, Karolina Sztandar-Sztanderska, Katarzyna Szymielewicz. 2015. Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decison Making. Warsaw: Fundacja Panoptykon. https://panoptykon.org/biblio/profiling-unemployed-poland-social-and-political-implications-algorithmic-decison-making. Dostęp 11.05.2018.

O'Neil, Cathy. 2017. *Broń matematycznej zagłady: jak algorytmy zwiększają nierówności i zagrażają demokracji*. Tłum. Marcin Zieliński. Warszawa: Wydawnictwo Naukowe PWN.

Pasquale, Frank. 2015. *The black box society: the secret algorithms that control money and information*. Cambridge: Harvard University Press.

Peña Gangadharan, Seeta, Jędrzej Niklas. 2019. Decentering Technology in Discourse on Discrimination. *Information, Communication & Society*, 22, 7: 882–99. DOI: 10.1080/1369118X.2019.1593484.

Piwowar, Kuba. 2019. Uprzedzenia w algorytmach. *Humanizacja Pracy*, 3, 297: 35–51.

Revelle, William. 2020. *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 2.0.8. Evanston, Illinois: Northwestern University. https://CRAN.R-project.org/package=psych.

Rieke, Aaron, Miranda Bogen, David G. Robinson. 2018. Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods. Upturn, Omidyar Network. https://omidyar.com/wp-content/uploads/2020/09/Public-Scrutiny-of-Automated-Decisions.pdf.

Sijtsma, Klaas. 2009. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, 74, 1: 107–20. DOI:10.1007/s11336-008-9101-0.

Steinberg, Lynne, David Thissen. 1996. Uses of Item Response Theory and the Testlet Concept in the Measurement of Psychopathology. *Psychological Methods*, 1, 1: 81–97. DOI:10.1037/1082-989X.1.1.81.

Sułek, Antoni. 2002. *Ogród metodologii socjologicznej*. Warszawa: Wydawnictwo Naukowe Scholar.

Sztandar-Sztanderska, Karolina. 2010. Poland: Recent Trends in Active Labour Market Policy. Internal report. Eurofund.

Sztandar-Sztanderska, Karolina. 2013. Nie zrzucajmy całej winy na nieefektywne urzędy pracy i fikcyjnych bezrobotnych. Uwagi o niepożądanych konsekwencjach nowelizacji Ustawy o promocji zatrudnienia i instytucjach rynku pracy. Warszawa: EAPN Polska. http://www.eapn.org.pl/eapn/uploads/2013/10/Ekspertyza_3.pdf.

Sztandar-Sztanderska, Karolina. 2016. *Obywatel spotyka państwo: o urzędach pracy jako biurokracji pierwszego kontaktu*. Warszawa: Wydawnictwo Naukowe Scholar.

Sztandar-Sztanderska, Karolina, Marianna Zieleńska. 2018. Changing social citizenship through information technology. *Social Work & Society, an International Online Journal*, 16, 2: https://ejournals.bib.uni-wuppertal.de/index.php/sws/article/view/566/1138.

Sztandar-Sztanderska, Karolina. 2020. What Makes an Ideal Unemployed Person? Values and Norms Encapsulated in a Computerized Profiling Tool. *Social Work & Society. International Online Journal*. https://www.socwork.net/sws/article/view/617/1210.

Theiss, Maria. 2018. *Lokalne obywatelstwo społeczne w polityce społecznej: przykład wychowania przedszkolnego*. Warszawa: Wydawnictwo Naukowe Scholar.

Wedel, Janine R. 2014. *Unaccountable: how elite power brokers corrupt our finances, freedom, and security*. First Pegasus Books cloth edition. New York: Pegasus Books.

Wiśniewski, Zenon, Monika Wojdyło-Preisner. 2013. *Profilowanie bezrobotnych wymagających szczególnego wsparcia na lokalnym rynku pracy*. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Wiśniewski, Zenon, Monika Wojdyło-Preisner. 2014. *Diagnozowanie stopnia zagrożenia długotrwałym bezrobociem: teoria i praktyka : poradnik profilowania bezrobotnych na lokalnym rynku pracy*. Warszawa: Ministerstwo Pracy i Polityki Społecznej : Centrum Rozwoju Zasobów Ludzkich.

Wiśniewski, Zenon, Monika Wojdyło-Preisner. 2015. Profilowanie bezrobotnych w Polsce i Niemczech. *Polityka Społeczna*, 2, 42: 22–27.

Wojdyło-Preisner, Monika. 2009. *Profilowanie bezrobotnych jako metoda przeciwdziałania długookresowemu bezrobociu*. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Zieleńska, Marianna. 2015. *Mechanizmy reprodukcji i zmiany w systemie administracji publicznej na przykładzie wdrażania otwartej metody koordynacji*. Warszawa: Wydawnictwo Naukowe Scholar.

Zieleńska, Marianna, Magdalena Tomasik. 2010. Diagnoza obecnego stanu – osoby niepełnosprawne w instytucjach aktywizacji społecznej i zawodowej. W: I. Wóycicka, red. *Skuteczność lokalnego systemu wsparcia na rzecz integracji społecznej i zawodowej osób niepełnosprawnych*. Warszawa: Instytut Badań nad Gospodarką Rynkową, 29–60.

Zouridis, Stavros, Mark Bovens, Marlies Van Eck. 2020. Automated Discretion. W: T. Evans, P. Hupe, eds. *Discretion and the Quest for Controlled Freedom*, Palgrave Macmillan, 313–229.

Zweig, Katharina A., Georg Wenzelburger, Tobias D. Krafft. 2018. On chances and risks of security related algorithmic decision making systems. *European Journal for Security Research*, 3: 181–203. https://doi.org/10.1007/s41125-018-0031-2.