

G r a ż y n a O s i k a

Datafikacja – implikacje epistemologiczne

Słowa kluczowe: *datafikacja*, „*big data*”, „*data science*”, *eksploracja danych*, *dataizm*, *algorytmizacja*

Wstęp

Bycie w świecie, mówiąc kategoriami Heideggerowskimi, wymaga „rozumienia wiedzy”¹, która jest nam pomocna w zadamawianiu się w rzeczywistości. A „rozumieć wiedzę” to poddawać ją filozoficznemu namysłowi, który, na co zwraca uwagę Wojciech Chudy, „nie dokonuje się w izolacji”.

Ów sposób i styl intelektualnego działania jest zakorzeniony w przeszłości, kształtuje go też dialog ze współczesnością. Nurty w dziejach filozofii tworzone są poprzez skomplikowaną dialektykę oddziaływań, zapożyczeń i zapośredniczeń. Funkcje te żłobią w atmosferze intelektualnej epok głębokie i trwałe łożysko, którym płynie dominujący strumień interpretującej myśli. W strumieniu filozofowania mieszczą się różnorakie elementy, w tym pragmatyczne i pozapoznawcze. Jednak trzeba tam także dostrzec i wyłonić element metody i wartości poznawcze (...) kierujące biegiem strumienia (Chudy 1995: 18).

Ten typ aktywności w dużym stopniu jest nieunikniony – rozumienie świata należy do podstawowych impulsów dążeń człowieka (Maslow 2006), pozwala mu bowiem wyznaczać horyzont jego sprawczości. Natomiast

Grażyna Osika, Politechnika Śląska, Katedra Stosowanych Nauk Społecznych, ul. F. Roosevelta 26-28, 41-800 Zabrze; e-mail: grazyna.osika@polsl.pl, ORCID: 0000-0002-8729-1001.

¹ Wyrażenie to należy identyfikować z takimi obszarami ludzkiej refleksji jak „teoria wiedzy”, „ogólna metodologia nauk”, „naukoznawstwo” itp., czyli tam, gdzie wiedzę traktujemy jako przedmiot poznania.

intensywność tej aktywności nasila się szczególnie wtedy, kiedy ulegają rozmyciu kategorie prawomocności poznania, tak, jak to się stało w przypadku późnej nowoczesności, kiedy wiedza zaczęła przybierać charakter refleksyjny, tj. rozpoczął się proces permanentnej weryfikacji (Giddens 1991), skutkujący ostatecznie upłynnieniem wiedzy. Z kolei jej fragmentaryzacja, związana z dziedzinową specjalizacją (Furedi 2004; Elias 2008: 221), w coraz większym stopniu zaczęła utrudniać budowanie szerszych kontekstów – holistycznych wizji rzeczywistości – a w konsekwencji, wbrew przyrostowi informacji o świecie, stawał się on dla człowieka coraz mniej czytelny. I tu zarysowuje się pierwsze pytanie badawcze, a mianowicie: jakie metody i wartości poznawcze kierują współcześnie biegiem strumienia wiedzy?

Pomocną podpowiedź w tym względzie możemy uzyskać kierując uwagę na narzędzia, które coraz częściej są stosowane w pracach badawczych – w tym przypadku wyraźnie dostrzegalna jest tendencja wykorzystania cyfrowych technologii i ich potencjału gromadzenia oraz przetwarzania danych. Wraz z nią odradzają się nadzieje na „obiektywizację” procesów wiedzytwórczych; coraz śміiej mówi się także o narodzinach czwartego paradygmatu nauki (Mayer-Schönberger, Cukier 2013), a więc takiego, który jest oparty na *big data* i *data science*.

Ponieważ w tym nowym procesie wydobywania wiedzy znaczącą rolę odgrywają maszyny, warto pamiętać o poczynionych już wcześniej w historii rozpoznaniach – mowa tu przede wszystkim o Wernerze Heisenbergu, który na początku XX wieku sformułował „zasadę nieoznaczoności”, przestrzegając nas, że już sam akt obserwacji ma wpływ na analizowane układy rzeczywistości, a narzędzia badawcze zmieniają uzyskiwane efekty poznawcze. Regułę tę ujął nieco później bardzo sugestywnym stwierdzeniem: „ktoś, kto w pracy używa maszyn, sam pracuje jak maszyna” (Heisenberg 2004: 110)². Świadomość tego faktu musi nasuwać kolejne pytanie badawcze: jak stosowane narzędzia wpływają na uzyskiwane efekty poznawcze? I tu ponownie, warto odwołać się do wcześniejszych ustaleń – tym razem chodzi o diagnozę stawianą przez Jean-François Lyotarda pod koniec lat 70., kiedy wyłaniały się zręby technologii cyfrowej. W *Kondycji ponowoczesnej* stwierdzał:

² Wydaje się, że na obecnym etapie badań niemożliwe jest już utrzymanie instrumentalnej koncepcji techniki, tj. rozpoznawanie jej jako zbioru obcych bytów, „które sami powołałismo do istnienia, kierując się racjonalnymi przesłankami usprawniania działania, (...) wszelkie relacje, jakie nawiązujemy z nimi, definiujemy z pozycji nadrzędnej. W tym prostomyślnym spojrzeniu oczywiste jest, że mamy pełne panowanie nad sytuacją, bo to człowiek wyznacza cele i wartości, jakim służy dane rozwiązanie techniczne” (Osika 2017: 70). Bowiemy wiemy już, że tworząc technologię, zaczynamy wchodzić z nią w interakcję, która zawsze ma charakter dwustronny (Freenberg 2002; Freenberg 2005; Dusek 2011; Bińczyk 2012).

skutki (...) technologicznych przekształceń wydają się mieć dla wiedzy ogromne znaczenie. (...) Wraz z hegemonią informatyki narzuca się (...) pewna logika, a także zestaw instrukcji odnoszących się do wypowiedzi uznawanych za należne do wiedzy (Lyotard 1997: 27).

Uwaga Lyotarda pozwala nam ostatecznie doprecyzować, o jaki rodzaj narzędzi chodzi, tj. postawić pytanie: jak technologia cyfrowa wpływa na to, co uznajemy za wiedzę?

Głównym celem dalszych rozważań jest opis zjawisk oraz zbadanie zależności pozwalających sformułować odpowiedzi na zadane wyżej pytania. Przyjmuje się, że nasilające się tendencje datafikacyjne mogą spowodować ukonstytuowanie się datacentrycznego sposobu postrzegania wszystkich aspektów rzeczywistości, czyniąc z danych oraz metod ich przetwarzania rodzaj wyższej instancji, kształtującej ludzki sposób myślenia o świecie. Badania będą miały charakter teoretyczny; zostaną przeanalizowane proces datafikacji oraz pojęcie *data science*, czyli to, co ujmuje się jako czwarty paradygmat nauki. Zostaną także sformułowane obszary niepewności, które powinny skupiać szczególną uwagę, bo „rozumieć wiedzę” to nade wszystko znać ograniczenia naszego sposobu jej „konstruowania”³.

Datafikacja

Kenneth Neil Cukier i Victor Mayer-Schoenberger w artykule *The Rise of Big Data* (2013), próbując uchwycić specyfikę trendu ujawniającego się w badaniach, zwracają uwagę, że obecnie dochodzi do „przekształcenia w dane wielu aspektów świata, które nigdy wcześniej nie zostały określone ilościowo”, i nazywają ten proces „datafikacją” (Cukier, Mayer-Schoenberger 2013). Można zatem powiedzieć, że pojęcie to narodziło się niejako naturalnie, z obserwacji efektów, jakie przyniosła technologiczna rewolucja tzw. wielkich danych (*Big Data*, BD), na którą składa się możliwości gromadzenia, przechowywania, przetwarzania i korelowania bardzo dużych zbiorów danych bez konieczności dbania o ich „czystość”. Autorzy rozwinęli to pojęcie w książce *Big Data: A Revolution That Will Transform How We Live, Work, and Think*: datafikacja to „przedstawianie określonego zjawiska w skwantyfikowanej formie, która może być ujmowana w tabelach i analizowana” (Mayer-Schönberger, Cukier 2013: 96).

³ Wyrażenie to wzięto w cudzysłów, żeby zdystansować się od ewentualnych sporów ontologicznych; nie jest ono także deklaracją światopoglądową autorki, jest raczej efektem przyjętej perspektywy badawczej, w której chodzi o przeanalizowanie zasadności pewnych założeń, i niczym więcej. Jest „testem” pewnej opcji, którą można lub powinno się mieć na uwadze.

Jednakże ich zdaniem, nie powinno się utożsamiać datafikacji z digitalizacją; datafikację należy rozpatrywać jako kontynuację starożytnych dążeń człowieka do „matematyzowania świata”, czyli poddawania go kwantyfikacji dzięki mierzeniu, rejestrowaniu, analizowaniu kategoriami przeliczalnymi, natomiast digitalizacja jest tym, co pozwoliło nadać tym procesom zupełnie nową jakość, przenosząc ją na wyższy poziom zaawansowania (Mayer-Schönberger, Cukier 2013: 97; Jałochowski 2016: 14–18). Od sformułowania pomysłu maszyny Turinga, jak wskazuje Yuval Noah Harari, informatycy rozwijali swoje umiejętności w zakresie projektowania coraz bardziej skomplikowanych elektronicznych algorytmów (Harari 2018: 467) i to stworzyło technologiczne zaplecze umożliwiające dokonywanie obliczeń na wcześniej niespotykaną skalę. W takim sensie digitalizację można uznać za drugi istotny czynnik składający się na obecne „matematycznie zorientowane” poznawanie świata; z jednej strony chcemy kwantyfikować, bo od zarania wieków chcieliśmy, ale teraz dysponujemy technologią, która nam na to pozwala, która wręcz nachyla naszą interakcję ze światem w stronę kategorii ilościowych (Manovich 2001: 27–30; Osika 2015: 72–74; Szpunar 2019: 11–22).

Istnieje jeszcze jeden aspekt wpływający na obecny zasięg procesów datafikacji – jest to uświadomiony wraz z upowszechnianiem cyfrowych mediów i powstaniem „internetu rzeczy” (IoT) potencjał generowania „śladów cyfrowych” (*digital footprints*) (van Dijck 2014; O’Neil 2016; Schutt, O’Neil 2014; Rudder 2014; Jurgenson 2014; Surma 2017; Jones 2019), począwszy od Web 1.0 po każdą kolejną generację poszerzającą ten zakres (van Dijck 2014: 202). Rozwój technologii informacyjnych stopniowo uświadamiał, że wszelka aktywność medialna pozostawia po sobie znak, rejestrowany „w ciszy”, „sekunda po sekundzie na wielkich farmach serwerów” (Surma 2017: 13). W „śladach” tych dostrzeżono „dane”, czyli „informacje”, „fakty” lub „liczby” (*Cambridge Dictionary*), pozyskiwane także z metadanych, które po przeanalizowaniu mogą stać się cennym wsparciem w procesach decyzyjnych. Początkowo istotny był sam fakt generowania dużej ilości danych, także ze względów czysto pragmatycznych, tj. „fizycznej” konieczności ich obsługi; z czasem zaczęto z nich robić bardziej konkretny użytek.

Podsumowując, gdyby chcieć uchwycić, co stoi u podstaw rodzącej się logiki datafikacji, należałoby wymienić charakterystyczną dla człowieka kwantyfikacyjną tendencję ujmowania rzeczywistości, cyfrowy potencjał technologiczny wraz z całym zapleczem obliczeniowym oraz bardzo proste spostrzeżenie, że sam fakt korzystania z cyfrowej technologii jest źródłem danych i że proces ten można próbować przenieść na inne aspekty życia, takie jak emocje, relacje, zwyczaje, długość czasu poświęcana na kontakty, sposób czytania strony itp., aż po „wszystko”, czyli $N = \text{WSZYSTKO}$ (Mayer-Schönberger, Cukier 2013).

Trafnie ten kwantyfikacyjny apetyt ujmuje José van Dijck – „duże, nigdy wcześniej niedostępne, nieświadome zbiory pozostawionych danych otwierają nowe widoki” (van Dijck 2014: 202). Dane uzyskiwane z różnych form śladów cyfrowych korelowane ze sobą zaczynają rozbudzać nadzieje na ilościowe ujęcie stanów emocjonalnych, nastrojów społecznych itp., jako „czujniki” wydarzeń odbywających się w czasie rzeczywistym lub „detektory uczuć” (van Dijck 2014: 199). Powstaje także nowy zawód – „górnika danych”, którego zadaniem jest odnajdywanie prawidłowości i przewidywanie tendencji dzięki zautomatyzowanym procesom analizy. Kluczowy w tym względzie staje się dostęp do odpowiednich zbiorów danych na temat ludzkich zachowań (Szeliga 2017: xix) oraz ich pogrupowanie, uporządkowanie we wzorce lub znaczące skorelowanie (McIlwraith, Marmanis, Babenko 2017: 41), tak aby móc je wykorzystać w sposób zoptymalizowany w przyszłości, tj. na ich podstawie dokonać prognozy i podjąć dostosowaną do konkretnych celów decyzję, także w czasie rzeczywistym. Adekwatnym przykładem tego typu działań są usługi, jakie wykonała Cambridge Analytica dla sztabu zwolenników Brexitu oraz w kampanii prezydenckiej Donalda Trumpa w 2016 roku, kiedy na podstawie 5000 punktów informacyjnych dotyczących poszczególnych osób, pozyskanych na platformach mediów społecznościowych, zastosowano psychografię, dzięki której udało się wyodrębnić osoby łatwiej podlegające wpływowi i wobec nich użyto odpowiednio dawkowanych, spersonalizowanych przekazów perswazyjnych, przesądzając w ten sposób o efektach głosowania.

Jak widać, z chwilą skwantyfikowania zjawisk, czyli ujęcia ich w formie danych, tj. w wyniku ich datafikacji, możemy przekształcać je i wydobywać z nich nowe wartości (O’Neil, Schutt 2014: 6), a także generować dla nich nowe zastosowania (Sivarajah i in. 2017). Pozostaje mieć nadzieję, że będą one służyły powszechnemu dobru społecznemu.

Należy jednakże pamiętać, że kwantyfikacja zjawisk jest tylko początkiem dochodzenia do wiedzy, która musi zostać wydobyta z tych danych dzięki zastosowaniu konkretnych metod określanych mianem *data science* (DS).

***Data science* jako metoda naukowa – nowy paradygmat naukowy**

Jak diagnozuje Hong Liu:

w XXI wieku mamy do czynienia z dużą ilością danych, które są obiektem badań tzw. metodami *data mining*, a poszukiwanie prawidłowości tych badań pozwoliło wyodrębnić nową dyscyplinę naukową określaną jako *data science*. Obecnie metoda ta stanowi uzupełnienie tradycyjnych metod naukowych opartych na hipotezach, jednak zmiany ewoluują w stronę przekształcenia paradygmatu naukowego z redukcjonistycznych

uproszczeń w złożoną systematyczną naukę, znacznie zmieniając nie tylko nauki ścisłe, ale również wywierając wpływ na technologiczną, nauki społeczne i ekonomiczne (Liu 2014: 61).

Data science lub *eScience*, a obok nich *data scientists* są zatem stosunkowo nowymi pojęciami w nauce, z którymi identyfikuje się zarówno pewne obszary badawcze, jak i metodę ich prowadzenia, a także specjalistów z tych zakresów. Cathy O’Neil i Rachel Schutt we wprowadzeniu do książki *Doing Data Science* uprzedzają, że są to pojęcia słabo zdefiniowane, bez wyraźnie wyznaczonych pól denotacyjnych (O’Neil, Schutt 2014: 1–2). Niemniej jednak, znając te ograniczenia, konieczne jest przynajmniej szkieletowe wskazanie obszaru badawczego oraz specyfiki wykorzystywanej metody.

Pojęcie *data science* po raz pierwszy wykorzystał Peter Naura w 1960 roku do określenia metod „automatycznego przetwarzania danych przy użyciu komputerów” (Szeliga 2017: xix). Od tego czasu funkcjonuje ono w środowiskach o proveniencji statystycznej w odniesieniu do „uczenia się” zarówno ludzi, jak i maszyn, dokonującego się na podstawie danych. Tym, co odróżnia je od tradycyjnych, statystycznych metod badawczych, jest bazowanie wyłącznie na danych cyfrowych (Szeliga 2017: xix). Celem jest automatyczne podejmowanie decyzji, prognozowanie i tworzenie wiedzy (Manovich 2015: 21).

Data science możemy więc zdefiniować jako interdyscyplinarny obszar badawczy dotyczący procesów i systemów naukowych służących do pozyskiwania wiedzy z danych cyfrowych lub dokładniej – jest to szeroka dziedzina, która odnosi się do złożonych procesów, teorii, koncepcji, narzędzi i technologii, które umożliwiają przegląd, analizę i ekstrakcję informacji oraz wiedzy z surowych danych (Dichev, Dicheva 2017: 2151). Dziedzina ta sytuuje się na pograniczu wiedzy matematycznej, statystycznej (Olhede, Wolfe 2018), uczenia maszynowego, umiejętności hakerskich oraz określonej wiedzy eksperckiej, której dane dotyczą (O’Neil, Schutt 2014: 7). Można mówić także o charakterystycznych umiejętnościach *data scientists*, które obejmują obrabianie danych, czyli ich eksplorację oraz wizualizację wyników (Yau 2009). Natomiast do nowych kompetencji zalicza się translację danych lub data-nautykę, zajmującą się obserwacją „przestrzeni” danych oraz wyjaśnianiem wydobywanych z danych znaczeń w celu lepszego ich wykorzystania (Jesionek 2018).

Charakterystyczne dla procesu badawczego DS jest wychodzenie od zanieczyszczonych danych, przechodzenie do modeli, a następnie wyciąganie wniosków. Natomiast w tradycyjnym podejściu badanie rozpoczyna się od postawienia hipotezy, następnie weryfikuje się jej zasadność i formułuje wnioski (Szeliga 2017: 12–14). Możliwość pominięcia etapu związanego ze stawianiem hipotezy, a w związku z tym odnoszenie się do określonej teorii, wyartykułował Chris Anderson w kontrowersyjnym artykule *The End of*

Theory: The Data Deluge Makes the Scientific Method Obsolete (2008), w którym przekonywał:

petabajty danych pozwalają na stwierdzenie: „korelacja jest wystarczająca” (...). Możemy analizować dane bez hipotez. Możemy wrzucić liczby do największych klastrów obliczeniowych, jakie kiedykolwiek widział świat, i pozwolić algorytmom odszukać wzorce tam, gdzie nauka nie może ich odnaleźć (Anderson 2008).

Postrzeganie datafikacji i *big data* jako rdzenia *data science* stoi u podstaw tzw. czwartego, obliczeniowego paradygmatu naukowego. Zdaniem Jima Graya, jego istotą jest ujednoczenie teorii dzięki eksperymentom i symulacjom dokonywanym w oparciu o cyfrowe technologie informatyczne (Gray 2009). Datafikacja pozwala dostarczać dużej ilości materiału badawczego, a cyfrowe algorytmy odślaniają związki istniejące między nimi, czyli dokonują tzw. eksploatawania rzeczywistości (Mayer-Schönberger, Cukier 2013; Hey, Tansley, Tolle 2009; Bendyk 2015: 111–115), dzięki któremu możemy dochodzić do rozumienia.

Rob Kichen w czwartym paradygmacie naukowym rozpoznaje odrodzenie empirycznej epistemologii, jednak zrywającej z podejściem dedukcyjnym. Jak bowiem wynika z założeń tego paradygmatu, *big data* dzięki datafikacji pozwala ująć całą rzeczywistość, nie wymaga tworzenia teorii *a priori*, modeli czy hipotez, a dzięki zastosowaniu metod analizy danych badanie jest agnostyczne, tj. „dane mogą mówić same za siebie, wolne od ludzkich uprzedzeń” (Kichen 2014: 4). Odnajdywane znaczące korelacje między danymi są z natury prawdziwe, dlatego uzyskiwana w ten sposób wiedza może wykraczać „poza kontekst lub wiedzę specyficzną dla danej dziedziny, a zatem może być interpretowana przez każdego, kto potrafi zdekodować statystykę lub wizualizację danych” (Kichen 2014: 4).

W procesie badania danych, czyli wydobywania wiedzy, możemy wyodrębnić kilka kluczowych etapów. Należy do nich przede wszystkim pozyskiwanie surowych danych, których źródłem jest wybrany aspekt rzeczywistości; następnie dane są obrabiane, czyszczone i poddawane eksploracyjnej analizie (*data mining*), która poprzedza zastosowanie zautomatyzowanych narzędzi, takich jak uczenie maszynowe, algorytmy, modele statystyczne (O’Neil, Schutt 2014: 41). Po przetworzeniu danych wyniki są wizualizowane i jako uzyskana wiedza mogą pomóc w podejmowaniu decyzji (analizy predykcyjne) lub służyć funkcjom wyjaśniającym (analizy deskryptywne) (Szeliga 2017: 12).

Za kluczową dla eksperymentów z zakresu *data science* uznaje się pionierską metodykę „odkrywania wiedzy” (*knowledge discovery*; Fayyad, Piatetsky-Shapiro, Smyth 1996), metodykę CRISP-DM (*Cross Industry Standard Process for Data Mining*) oraz ujęcie zaproponowane w dokumencie *Challenges and Opportunities with Big Data* (por. Szeliga 2017: 2–3). We wszystkich wymienionych podejściach zakłada się, że centralnym elementem

procesu są dane, z których następnie, po przekształceniu i modelowaniu, wydobywa się zawartą w nich wiedzę.

Z punktu widzenia obiektywizacji warunków i efektów poznania propozycje zawarte w czwartym, obliczeniowym paradygmacie naukowym wyglądają obiecująco, wydaje się jednak, że zanim „kupimy” to rozwiązanie z całym bogactwem inwentarza, warto poddać je analizie krytycznej, mogącej odsłonić ewentualne obszary niepewności. Jest to szczególnie istotne ze względu na technologiczny *bias* badaczy *data science*, których uwaga kieruje się głównie w stronę technicznego zaplecza eksperymentu, czyli ku odpowiedzi na pytanie „jak to zrobić?” Natomiast z filozoficznej perspektywy ważniejsze jest pytanie „czy te narzędzia mają szansę przybliżyć nas do lepszego rozumienia rzeczywistości?” lub „jaki jest świat, którego dotyczy ta wiedza?”

Datafikacja – obszary niepewności

Jeżeli zgodzimy się, że metody *data science* stają się obecnie znaczącym elementem głównego nurtu poznawczego, a jak wykazano powyżej, istnieją ku temu przesłanki, wówczas zasadne wydaje się rozpatrzenie, jak metody te mogą wpływać na to, co uznajemy za wiedzę. Poniżej zostaną opisane tylko te aspekty, które mogą generować najistotniejsze obszary niepewności.

Po pierwsze, kiedy zaczynamy uznawać *data science* za metodę naukową, musimy zadać pytanie o granice datafikacji, tj. czy możliwe jest próbkowanie „wszystkiego”? Czy możemy przyjąć, że wszystkie procesy rzeczywistego świata zostawiają „ślady”, czyli dane, które mieszczą się w technologicznie usankcjonowanych metodach pozyskiwania danych? W tym przypadku musimy mieć świadomość zagrożenia, które Roy Harris określał mianem efektu „ślepej plamki” (Harris 2014: 25). Rozpoznany przez Harrisa efekt polega, z grubsza rzecz biorąc, na braku świadomości, jak fakt posługiwania się technologią wpływa na sposób, w jaki postrzegamy rzeczywistość. Musimy zatem liczyć się z tym, że jeżeli jakiś aspekt rzeczywistości nie będzie podlegał cyfrowej kwantyfikacji, to zostanie z niej usunięty, ulegnie redukcji. Powstaje pytanie: czy możemy sobie pozwolić na taką nonszalancję wobec rzeczywistości? I nie chodzi tu wyłącznie o konieczność akceptacji *blind spotów* – z tym ludzkość zdążyła się już oswoić – ale o to, czy kiedy wiemy, że istnieją obszary niedatyfikowalne, możemy je pomijać i twierdzić, że nasze poznanie jest prawomocne? Nawet jeżeli uznamy te granice, datacentryzm pozostaje faktem.

Kolejny problem dotyczy ukrytego błędu kategorialnego. Datafikacja wbrew pozorom nie jest związana wyłącznie z metodami badawczymi, lecz zawiera w sobie zaszytą teorię ontologiczną (Crawford 2014: 1664), zgodnie

z którą „wszechświat składa się z przepływu danych, a wartość każdego zjawiska czy bytu określa jego wkład w przetwarzanie danych” (Harari 2018: 467). Pogląd ten jest określany mianem „dataizmu” (Brooks 2013; van Dijck 2014; Harari 2018). Zdaniem Davida Brooksa, twórcy tego terminu, „dataizm” możemy uznać za wschodzącą filozofię, zgodnie z którą „dane są przejrzyste i wiarygodną soczewką”, pozwalającą nam „filtrować” rzeczywistość i umożliwiającą dzięki temu przepowiadanie przyszłości (Brooks 2013). Dataizm zakłada, że wszystkie aspekty rzeczywistości podlegają obiektywnej kwantyfikacji (van Dijck 2014: 201) oraz że możliwe jest odkrycie naczelnego algorytmu, który pozwoli ująć przeszłą, obecną i przyszłą wiedzę w całość i odsłonić ZASADĘ świata (Domingos 2016).

Dla polityków, przedsiębiorców i zwykłych konsumentów dataizm jest czymś, co daje przełomowe technologie i ogromne możliwości. Uczonym i intelektualistom (...) obiecuje (...) stworzenie jednej wszechobjemującej teorii, łączącej wszystkie dyscypliny naukowe: od muzykologii przez ekonomię po biologię. Według dataizmu *V Symfonia* Beethovena, bańka giełdowa i wirus grypy to tylko trzy wzory przepływu danych, trzy prawidłowości, które można analizować z użyciem tych samych podstawowych pojęć i narzędzi (Harari 2018: 467–468).

Dataizm zakłada, że matematyka nie tyle jest skutecznym językiem opisu rzeczywistości (Heller 2016: 66), ile raczej stanowi jej własność. Oczywiście takie podejście pozwala znieść problem istnienia możliwych niefunkcyjnych aspektów świata – po prostu przyjmuje się, że składa się on z danych, które możemy cyfrowo kwantyfikować, porządkować algorytmami, czyniąc w ten sposób świat inteligibilnym. Nadal jednak nie umiemy rozstrzygnąć, czy model ten jest adekwatny do rzeczywistości – czy wiernie ją „relacjonuje”.

Wydaje się również, że wbrew składanym obietnicom, mamy w tym przypadku do czynienia raczej ze „słabą” wersją obiektywizmu – dotyczy to zarówno samych danych, jak i stosowanych modeli. Jeżeli nie ulegniemy czarowi dataizmu, łatwo dostrzeżemy zjawisko, na które zwraca uwagę Cathy O’Neil – rozumienie danych jest procesem subiektywnym. Subiektywność ta wyłania się już na samym początku, kiedy to, w jakimś stopniu arbitralnie, są podejmowane decyzje na temat tego, co uznaje się za dane. Kolejna pułapka dotyczy etapu eksploracji danych, gdy są wybierane modele i algorytmy, którymi dane zostaną przebadane. Należy przy tym pamiętać, że:

żaden model nie może uwzględnić wszystkich komplikacji realnego świata (...). Tworząc model, dokonujemy wyboru tego, co wydaje się (...) wystarczająco ważne, by to uwzględnić. Upraszczamy świat. (...) Przyjmujemy, że model będzie realizował tylko jedno zadanie, oraz godzimy się z tym, że od czasu do czasu będzie zachowywał się jak nierozumna maszyna, skazana na olbrzymie luki informacyjne (...). Luki informacyjne odzwierciedlają oceny i priorytety ich twórców (...). Modele są opiniami napisanymi językiem matematyki (O’Neil 2017: 46–48).

Można mówić o istnieniu ustalonych konwencji interpretacji danych, ale konwencje te mogą być nieadekwatne, dlatego należy analizować zarówno to, co poddajemy analizie, tj. co uznajemy za dane, ale także sposób ich interpretacji (Jones 2019: 6–12; Gitelman, Jackson 2013). Robert D. Galliers i in. zwracają uwagę na określone wzory użycia datafikacji i ich niezamierzone konsekwencje, zaliczając do nich: dekontekstualizację danych, czyli wykorzystanie ich w innych kontekstach niż te, dla których zostały zebrane i przetworzone, rekombinację danych, polegającą na tworzeniu nowych danych dzięki dekontekstualizacji i korelacji, stosowanie danych ilościowych jako miar zastępczych dla złożonych zjawisk, strategiczne i selektywne wykorzystywanie danych do realizacji partykularnych interesów, legitymację wnioskowanych informacji w oparciu o legitymację danych pierwotnych, pomijanie roli ekspertów w eksperymentach itp. (Galliers i in. 2017: 187–188).

Data scientists często mówią także o tzw. „torturowaniu danych”⁴, kiedy nie tyle zgodnie z założeniem czwartego paradygmatu naukowego „podążamy za danymi”, ile tak długo poddajemy je obróbce, aż wynik stanie się zgodny z naszymi oczekiwaniami. Trudność, na którą często zwraca się uwagę, stanowi także tzw. *overfitting*, czyli podatność na „przeuczenie” modelu – jest to sytuacja, w której po zastosowaniu pewnego zbioru danych do estymowania modelu otrzymany model nie odzwierciedla rzeczywistości wskutek nadmiernego dopasowania (Griffiths, Christian 2018: 222–226; O’Neil, Schutt 2014: 34), a wtedy mimo przeprowadzonego eksperymentu uzyskiwane wnioski są nieprzydatne ani w kategoriach deskryptywnych, ani predyktywnych, ponieważ w jakimś sensie są już „o świecie minionym”.

Powstaje pytanie, czy taka nieadekwatność modelu zostanie dostrzeżona, bowiem uzyskiwana w ten sposób wiedza odznacza się silną sugestywnością prawomocności poznania dzięki mechanizmowi psychologicznemu, który przez analogię do pewnej innej heurystyki (tzw. społecznego dowodu słuszności) możemy nazwać „technologicznym dowodem słuszności” (Osika 2019: 190). Z grubsza rzecz biorąc, w sytuacjach poznawczej niepewności mamy tendencje do posiłkowania się wyższą instancją, która pozwala nam opanować dysonans poznawczy. W przypadku „dowodu społecznego” wierzymy w nieomyślność grupy – „skoro wszyscy tak to widzą, to musi to być prawda”, natomiast w przypadku dowodu technologicznego chodziłoby o powierzenie myślenia algorytmom – „skoro takie są obliczenia...” Wobec „obiektywnych” efektów eksploracji danych możemy mieć trudności z krytyczną refleksją. Ostatnia z formułowanych w ramach niniejszych rozważań wątpliwości dotyczy merytorycznej zawartości „wydobywanej” wiedzy i jej technologicz-

⁴ Określenie to przypisuje się Ronaldowi Coase, ekonomicznemu nobliście z 1991 roku, który podobno zwykł mawiać: „if you torture the data long enough it will confess”, uzmysławiając w ten sposób stopień obiektywizmu metod statystycznych (Tullock 2001).

nego ukierunkowania bardziej w stronę pytań „co?” i „jak?” niż „dlaczego?”. Skutkuje to pojmowaniem wiedzy jako informacji o faktach bez konieczności rozumienia ich sensu.

Brak świadomości istnienia opisanych powyżej obszarów niepewności wobec dokonujących się procesów datafikacji może prowadzić do zbyt łatwowiernego stosowania metod *data science*. Już teraz datacentryczne ujmowanie wielu aspektów rzeczywistości wydaje się niekwestionowalne. Tylko jednak uzmysłowienie sobie ograniczeń tych metod może nas uchronić przed kuszącym delegowaniem myślenia na maszyny.

Podsumowanie

Celem naszych rozważań był opis zjawisk związanych z procesem datafikacji oraz analiza zależności pozwalających odpowiedzieć na wymienione we wstępie pytania badawcze.

Jak się wydaje, musimy, po pierwsze, mieć świadomość, że datafikacja staje się coraz szybciej głównym nurtem badawczym. Nawet jeżeli w tej chwili dotyczy głównie nauk ścisłych, coraz więcej mówi się o tzw. humanistyce cyfrowej (Bomba 2013). Po drugie, musimy pamiętać o wpływie, jaki datafikacja może mieć na treść tego, co poznajemy, i w tym sensie musimy rozumieć metody związane z czwartym paradygmatem naukowym. Po trzecie, musimy zdawać sobie sprawę, że upowszechnienie datafikacji dokonuje przemian w tym, co uznajemy za wiedzę, a nawet tym, co uznajemy za naukę, przenosząc akcenty na datafikowalność jej treści i kwantyfikowalność efektów.

Przeprowadzone refleksje, w których poddano analizie takie zagadnienia jak datafikacja, *data science*, czwarty paradygmat nauki oraz możliwe obszary niepewności, wydają się stanowić wystarczające przesłanki do potwierdzenia stawianej we wprowadzeniu tezy, tj. że nasilające się tendencje datafikacyjne mogą spowodować ukonstytuowanie się datacentrycznego sposobu postrzegania wszystkich aspektów rzeczywistości, czyniąc z danych oraz metod ich przetwarzania rodzaj „wyższej instancji”, kształtującej ludzki sposób myślenia o świecie. W tym kontekście Magdalena Szpunar mówi wręcz o narodzinach „kultury algorytmów”:

funkcjonujemy w rzeczywistości społecznej poddanej niezdolnemu dyktatowi policzalności wszystkiego. Wartość i znaczenie ma jedynie to, co jest lub może być mierzalne. Kwantyfikacja stała się naszym fetyszem, bez którego nie wyobrażamy sobie życia we współczesnym świecie. W coraz mniejszym stopniu pozytywnie ewaluujemy typowo ludzkie atrybuty, wypierane przez tryby technicznej, odhumanizowanej, nastawionej na skuteczność maszyny społecznej (Szpunar 2019: 11).

Nawet jeżeli te tendencje nie wydają się nam aż tak wyraźne, nie wolno nam pomijać ich znaczenia oraz zakresu oddziaływania na obecny kształt społeczeństwa. Konieczna jest nieustanna rekonstrukcja procesów wytwarzania, rejestracji i wykorzystania danych (Jones 2019: 15). Pozwoli ona lepiej rozumieć człowieka w jego aktualności, a także odpowiedzieć na pytanie, jaki jest świat, o którym człowiek gromadzi swoją wiedzę – i czy jest to nadal świat jego dawnych wartości?

Bibliografia

- Anderson Ch. (2008), *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, „Wired”, <https://www.wired.com/2008/06/pb-theory/> [20.02.2020].
- Bendyk E. (2015), *Koniec Teorii*, „Niezbędnik Inteligent” 2, s. 111–115.
- Bińczyk E. (2012), *Technonauka w społeczeństwie ryzyka*, Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
- Bomba R. (2013), *Narzędzia cyfrowe jako wyznacznik nowego paradygmatu badań humanistycznych*, w: A. Radomski, R. Bomba (red.), *Zwrot cyfrowy w humanistyce. Internet / nowe media / kultura 2.0*, Lublin: E-naukowiec, s. 57–72.
- Brooks D. (2013), *The Philosophy of Data*, „The New York Times”, <https://www.nytimes.com/2013/02/05/opinion/brooks-the-philosophy-of-data.html> [03.03.2020].
- Cambridge Dictionary*, <https://dictionary.cambridge.org/pl/dictionary/english/data>, [21.01.2020].
- Chudy W. (1995), *Rozwój filozofowania a „pułapka refleksji”. Filozofia refleksji i próby jej przezwyciężenia*, Lublin: Wydawnictwo KUL.
- Crawford K., Miltner K., Gray M.L. (2014), *Critiquing Big Data: Politics, Ethics, Epistemology*, „International Journal of Communication” 8, s. 1663–1672, <https://ijoc.org/index.php/ijoc/article/viewFile/2167/1164>
- Cukier K.N., Mayer-Schoenberger V. (2013), *The Rise of Big Data*, <https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data>, „Foreign Affairs” [15.01.2020].
- Dichev Ch., Dicheva D. (2017), *Toward Data Science Literacy*, „Procedia Computer Science” 108C, s. 2151–2160.
- Domingos P. (2016), *Naczelny Algorytm*, przeł. R. Ociepa, Gliwice: Helion.
- Dusek V. (2011), *Wprowadzenie do filozofii techniki*, przeł. Z. Kasprzyk, Kraków: Wydawnictwo WAM.
- Fayyad U., Piatetsky-Shapiro G., Smyth P. (1996), *From Data Mining to Knowledge Discovery in Databases*, „AI Magazine” 17 (3), s. 37–54, 1230-Article-Text-1227-1-10-20080129.pdf [21.02.2020].
- Freenberg A. (2002), *Transforming Technology. Critical Theory Revisited*, Oxford – New York: Oxford University Press, https://monoskop.org/images/d/d8/Feenberg_Andrew_-_Transforming_Technology_A_Critical_Theory_Revisited.pdf [19.02.2020].
- Freenberg A. (2005), *Critical Theory of Technology: An Overview*, „Tailoring Biotechnologies” 1, <https://www.sfu.ca/~andrewf/books/critbio.pdf> [19.02.2020]

- Furedi F. (2004), *Where Have All the Intellectuals Gone? Confronting 21st Century Philistinism*, London – New York: Continuum International Publishing Group.
- Galliers R.D., Newell S., Shanks G., Topi H. (2017), *Datafication and its human, organizational and societal effects: The strategic opportunities and challenges of algorithmic decision-making*, „Journal of Strategic Information System” 26, s. 185–190.
- Giddens A. (1991), *Modernity and Self-Identity. Self and Society in the Late Modern Age*, Stanford: Stanford University Press.
- Gitelman L., Jackson V. (2013), *Introduction*, w: L. Gitelman (ed.), *“Raw Data” Is an Oxymoron*, Cambridge: MIT Press, s. 1–14.
- Gray J. (2009), *The Transformed Scientific Method*, w: T. Hey, S. Tansley, K. Tolle (eds.), *The Fourth Paradigm. Data-Intensive Scientific Discovery*, Redmond, Washington: Microsoft (e-book).
- Griffiths T., Christian B. (2018), *Algorytmy. Kiedy myśleć mniej*, przeł. D. Rossowski, Łódź: Feeria Science.
- Harari Y.N. (2018), *Homo deus*, przeł. M. Romanek, Warszawa: Wydawnictwo Literackie.
- Harris R. (2014), *Racjonalny umysł piśmienny*, przeł. M. Rakoczy, Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.
- Heller M. (2016), *Czy fizyka jest nauką humanistyczną?*, Kraków: Copernicus Center Press.
- Hey T., Tansley S., Tolle K. (eds.) (2009), *The Fourth Paradigm. Data-Intensive Scientific Discovery*, Redmond, Washington: Microsoft (e-book).
- Jałochowski K. (2016), *Mathemamorfozy*, „Niezbędnik Inteligenta” 1, s. 14–18.
- Jesionek R. (2018), *Translatorzy, datanauci i samuraje danych*, „Digital&More”, <https://digitalandmore.pl/translatorzy-datanauci-i-samuraje-danych/> [20.02.2020].
- Jones M. (2019), *What we talk about when we talk about (big) data*, „Journal of Strategic Information Systems” 28, s. 3–16.
- Jurgenson N. (2014), *View From Nowhere*, <https://thenewinquiry.com/view-from-nowhere/>
- Kichin R. (2014), *Big Data, new epistemologies and paradigm shifts*, „Big Data & Sociology”, https://people.cs.kuleuven.be/~bettina.berendt/teaching/VienaDH15/kitchin_2014.pdf
- Liu H. (2014), *Philosophical Reflection on Data*, „Procedia Computer Science” 30, s. 60–65.
- Liotard J.-F. (1997), *Kondycja ponowoczesna. Raport o stanie wiedzy*, przeł. M. Kowalska, J. Migasiński, Warszawa: Fundacja Aletheia.
- Manovich L. (2001), *The Language of New Media*, Cambridge: MIT Press.
- Manovich L. (2015), *Data Science and Digital Art History*, „DAH-Journal” 1, s. 13–35.
- Maslow A. (2006), *Motywacja i osobowość*, Warszawa: Wydawnictwo Naukowe PWN.
- Mayer-Schönberger V., Cukier K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, London: John Murray.
- Mellwraith D., Marmanis H., Babenko D. (2017), *Inteligentna sieć. Algorytmy przyszłości*, przeł. T. Walczak, Gliwice: Helion.

- McLuhan M. (2004), *Zrozumieć media. Przedłużenia człowieka*, przeł. N. Szczucka, Warszawa: Wydawnictwo Naukowo-Techniczne.
- Olhede S.C., Wolfe P.J. (2018), *The future of statistic and data science*, „Statistics and Probability Letters” 136, s. 46–50.
- O’Neil C. (2017), *Broń matematycznej zagłady*, przeł. M. Zieliński, Warszawa: PWN.
- O’Neil C., Schutt R. (2014), *Doing Data Science: Straight Talk from the Frontline*, Sebastopol: O’Reilly Media.
- Osika G. (2015), *Hiperczytanie jako struktura poznawcza*, „Lingua ac Communitas” 25, s. 63–78.
- Osika G. (2017), *Czekając na osobliwość – o modelach interpretacji techniki*, „Filosofija” 39/1, s. 65–78.
- Osika G. (2019), *Algorytmiczne narzędzia analizy zawartości mediów – metametodologiczne refleksje*, w: I. Hofman, D. Kępa-Figura (red.), *Współczesne media. Problemy i metody badań nad mediami*, Lublin: Wydawnictwo Uniwersyteckie UMCS, s. 181–194.
- Rudder Ch. (2014), *Dataclysm. Who We Are (When We Think No One’s Looking)*, New York: Crown.
- Sivarajah U., Kamal M.M., Irani Z., Weerakkody V. (2017), *Critical analysis of Big Data challenges and analytical methods*, „Journal of Business Research” 70, s. 263–286.
- Surma J. (2017), *Cyfryzacja życia w erze „big data”. Człowiek. Biznes. Państwo*, Warszawa: Wydawnictwo Naukowe PWN.
- Szeliga M. (2017), *„Data Science” i uczenie maszynowe*, Warszawa: PWN.
- Szpunar M. (2019), *Kultura algorytmów*, Kraków: Wydawnictwo ToC.
- Tulloch G. (2001), *A Comment on Daniel Klein’s „A Plea to Economists Who Favor Liberty”*, „Eastern Economic Journal” 27, nr 2, s. 203–207.
- van Dijck J. (2014), *Datafiction, dataism and dataveillance: Big Data between scientific paradigm and secular belief*, „Surveillance & Society” 12, nr 2, s. 197–208, <https://dare.uva.nl/search?identifier=eae3e10f-6cad-440b-8b0e-f623b9bdda9f> [19.02.2020].
- Yau N. (2009), *Rise of the Data Scientists*, <https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/> [20.02.2020].

G r a ż y n a O s i k a

Datafication – the epistemological implications

Keywords: *datafication, big data, data science, data mining, dataism, algorithmization*

A common observation of everyday life reveals the growing importance of data science methods, which are increasingly more and more important part of the mainstream of knowledge generation process. Digital technologies and their potential for data

collection and data processing have initiated the birth of the fourth paradigm of science, based on Big Data. Key to these transformations is datafication and data mining that allow the discovery of knowledge from contaminated data. The main purpose of the considerations presented here is to describe the phenomena that make up these processes and indicate their possible epistemological consequences. It has been assumed that increasing datafication tendencies may result in the formation of a data-centric perception of all aspects of reality, making data and the methods of their processing a kind of higher instance shaping human thinking about the world. This research is theoretical in nature. Such issues as the process of datafication and data science have been analyzed with a focus on the areas that raise doubts about the validity of this form of cognition.