Roczniki Kulturoznawcze
Tom XVI, numer 4 – 2025
https://doi.org/10.18290/rkult25164.2

JAN KREFT

SHADOW BANING: MIĘDZY „MINISTERSTWEM PRAWDY”, „TAJNĄ POLICJĄ” ALGORYTMICZNĄ A POZYSKIWANIEM WIEDZY O UŻYTKOWNIKACH

WPROWADZENIE

Przeróżnych opowieści o kontrowersjach, dyskusjach i reakcjach na temat shadow baningu jest bez liku. Ta opresyjna, ale i oczekiwana społecznie praktyka była i jest bowiem codziennym doświadczeniem „zwykłych” i „niezwykłych” użytkowników.

Pierwotnie shadow baning opisywał oszukańcze zawieszanie kont na forach internetowych: podczas gdy użytkownik był przekonany, że nadal może publikować swe treści i będą one widoczne dla innych, w rzeczywistości nikt ich nie widział. Obecnie w akademickiej dyskusji zwykle dotyczy usuwania z rankingów, ewentualnie obniżania rankingu, przy czym zachowany jest dostęp do treści, ale jest znacznie mniej widoczna za pośrednictwem wyszukiwania i nieobecna albo niemal nieobecna w rekomendacjach[1].

W szerszym ujęciu shadow baning to radykalne zmniejszanie widoczności treści i egzekwowanie enigmatycznych zasad działania platform przez blokowanie sugestii wyszukiwania, działań i zmniejszenie zaangażowania obserwujących. W praktyce shadow baning ma dotyczyć pojawiających się w internecie i niepożądanych treści dezinformacyjnych, clickbaitu i tzw. mowy nienawiści. „Walka”

Prof. dr hab. Jan Kreft – Politechnika Gdańska; adres do korespondencji: Gabriela Narutowicza 11/12, 80-222 Gdańsk; e-mail: jankreft@pg.edu.pl; ORCID: https://orcid.org/0000-0003-4129-8424.

[1] Kelley Cotter, „‘Shadowbanning is not a thing’: Black box gaslighting and the power to independently know and credibly critique algorithms”, *Information, Communication & Society* 26, nr 6 (2023):1226-1243.

z nimi może nie być ujawniana użytkownikom ze względu na koszt i złożoność tego procesu oraz nowość technologii[2].

Shadow baning jest zatem terminem nieprecyzyjnym i być może dlatego rzadko obecnym w relacjach medialnych, a tym bardziej w naukowej dyskusji na temat moderowania treści i zarządzania platformami[3]. Nie brakuje także głosów[4], że powinien być zastąpiony przez inny termin: „nieujawnioną moderację treści", ponieważ pozwala firmom platform mediów społecznościowych unikać odpowiedniego opisywania praktyk moderacyjnych, ułatwia aktorom politycznym wykorzystanie go do własnych celów, ponadto jego konspiracyjna natura umożliwia platformom używania go do opisywania moderacji, tym samym jeszcze bardziej spychając je w cień.

Tradycyjnie firmy platform odrzucają oskarżenia o stosowanie shadow baningu, choć niekiedy przyznają się do stosowania ograniczeń widoczności. Na uwagę zasługuje przegląd zasad obniżania rankingu. W przypadku Facebooka to zapis znany jako „Wytyczne dotyczące dystrybucji treści", czyli treści „na granicy", szkodliwe dezinformacje oraz „śmieci niskiej jakości" (clickbait, konkursy z nagrodami i linki do oszukańczych lub złośliwych witryn) oraz treści niskiej jakości, w tym nieoryginalne lub przerobione, jak również takie, których pochodzenie jest niejasne. W przypadku Instagrama mamy do czynienia z „kontrolą treści wrażliwych" (sygnalizowaną w 2021 roku). Jakie treści były poddane shadow baningowi, można było przekonać się pośrednio, ponieważ platforma pozwoliła, by użytkownicy mogli na nie „zezwolić", „ograniczyć (domyślnie)" lub „ograniczyć jeszcze bardziej" (chodziło o „treści seksualne, związane z bronią palną, narkotykami samookaleczenia i dezinformacje). YouTube skupiało się na treściach, które są bliskie, ale nie przekraczają granicy naruszania „wytycznych społeczności": filmów clickbaitowych z mylącymi tytułami i opisami, które otrzymały zbyt wiele podobnych rekomendacji, promujących fałszywe cudowne lekarstwo na poważną chorobę, twierdzących, że Ziemia jest płaska, lub

[2] Paddy Leerseen, „An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation", *Computer Law & Security Review* 48 (2023):105790.

[3] Shadow baning należy do pakietu strategii moderacyjnych, obok strategii dogmatycznych (użytkownicy są wyciszani, blokowani lub komentarze są usuwane bez podania powodu), autorytarno-interaktywnych (użytkownicy są informowani o zasadach moderacji i ich egzekwowaniu) i dyskursywno-interaktywnych (moderatorzy angażują się w dyskusje z użytkownikami). Zob. Andrea Stockinger, Svenja Schafer i Sophie Lechler, „Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools", *New Media & Society* (2023):14614448231190901.

[4] Gabriel Nicholas, „Sunsetting 'Shadowbanning'", *Yale Law School Information Society Project Platform Governance Terminologies Essay Series* (2023).

przedstawiających rażąco fałszywe twierdzenia dotyczące wydarzeń historycznych, takich jak atak na World Trade Centre 11 września[5].

Uogólnianie shadow baning dotyczy treści, które są na wyznaczonej przez komercyjne platformy „granicy" tego, co dozwolone z punktu widzenia społecznego dobrostanu, z drugiej strony platformy zdają się kierować przede wszystkim satysfakcją użytkowników jako konsumentów i unikać precyzyjnego identyfikowania tego, co wolno użytkownikowi, a czego – nie.

1. KONCEPCJA I METODA BADAWCZA

Kierując się potrzebą poznania shadow baningu jako niemarginalnego zjawiska na pograniczu dziedzin, w szczególności dyscyplin nauk społecznych i technicznych, wyznaczono cele badania: (1) nakreślenie krajobrazu badawczego poprzez identyfikację dominujących trendów badawczych i (2) syntezę kierunków przyszłych badań.

W badaniu zastosowano wspieraną maszynowo metodę systematycznego przeglądu literatury w zgodzie z ramami metodologicznymi autorstwa Marka Petticrew i Helen Roberts[6] oraz sformułowano następujące pytania badawcze: jaki jest stan badań nad shadow baningiem w odniesieniu do wybranych danych bibliometrycznych? Jaki nurt debaty wyłania się na przyszłość?

W pierwszym kroku przyjętej procedury zdefiniowano pięć terminów wyszukiwania. Na podstawie istniejącego dyskursu wybrano: „moderation", „shadow baning", „algorithm", „folk stories" i „algorithmic imagination" oraz „actor" (w rozszerzonej formie, obejmującej firmy medialne). Dane zebrano w okresie od 5 do 14 stycznia 2025 roku w ramach baz Scopus i Google Scholar – ich wybór wynika z dostępu do obszernej literatury naukowej z zakresu nauk społecznych, w tym także do nowych artykułów. Wyszukiwanie ograniczono do recenzowanych artykułów na temat shadow baningu opublikowanych w latach 2015–2024. Wcześniejsze badania nie były brane pod uwagę, ponieważ początkowe ustalenia mogły okazać się nieaktualne.

W ramach kilkustopniowego procesu wyszukiwania, po przeanalizowaniu wybranych głównych tematów, zidentyfikowano synonimy, terminy ogólne i terminy powiązane z terminami wyszukiwania (np. „folk stories – imagination algorithm"). Następnie przeprowadzono wyszukiwanie obejmujące wszystkie sześć

[5] Tarleton Gillespie, „Do not recommend? Reduction as a form of content moderation", *Social Media+ Society* 8, nr 3 (2022):20563051221117552.

[6] Mark Petticrew i Helen Roberts, *Systematic reviews in the social sciences: A practical guide* (Oxford: John Wiley & Sons, 2008).

terminów, co dało łącznie 146 wyników. W ostatnim kroku przeprowadzono dodatkowe wyszukiwania z czterema terminami wyszukiwania („shadow baning", „moderation", „folk stories", „imaginaries"). Następnie skonstruowano listę publikacji na podstawie tytułu, autora, informacji o publikacji i streszczenia każdej publikacji, przejrzano listę i zachowano tylko badania odpowiadające wybranemu przedziałowi czasowemu. W kolejnym kroku wyselekcjonowane zostały oryginalne, recenzowane artykuły dostępne w postaci pełnego tekstu. Ostatecznie, po zastosowaniu oprogramowania MAXQDA, które pozwoliło wykluczyć badania niemówiące wprost o shadow baningu, wyselekcjonowano 27 artykułów (załącznik 1), głównie w języku angielskim (także z tego powodu w artykule zachowano oryginalne angielskie terminy związane z przedmiotem badań). Ich analiza ujawniła główne, przedstawione poniżej, tematy badań, które pokrótce zaprezentowano. Zważywszy na ewolucyjną naturę pola badawczego, w zgodzie z drugim pytaniem badawczym zsyntezowano luki badawcze i wyodrębniono kluczowe ścieżki przyszłych badań.

2. GŁÓWNE NURTY DYSKUSJI

2.1. SHADOW BANING A MODERACJA

W środowisku cyfrowym dochodzi do licznych i nie zawsze uświadamianych przez użytkowników przejawów moderacji, czyli, ujmując definicyjne, mechanizmów zarządzania, algorytmicznych i przez człowieka, które strukturyzują uczestnictwo w społeczności w celu ułatwienia współpracy i zapobiegania nadużyciom[7]. Moderacją jest bowiem każda ingerencja, czy to wobec pojedynczego postu, czy wszystkich, każda klasyfikacja z wykorzystaniem uczenia maszynowego czy redukcja całej kategorii, jak w przypadku treści politycznych na Facebooku od końca ubiegłej dekady czy zmiany w reedukacji i popieraniu różnych treści na tej samej platformie w latach 2016–2020.

W praktyce najbardziej znanymi technikami moderacji są: całkowite zawieszenie, usunięcie lub tzw. deplatforming użytkowników lub treści[8]. Gdy platformy nie chcą usuwać treści, mają do wyboru, oprócz shadow baningu, inne środki[9], takie jak nakładanie barier wiekowych, blokowanie ze względu

[7] James Grimmelmann, „The virtues of moderation", *Yale Journal Law & Technics* 17 (2015):42.

[8] Richard Rogers, „Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media", *European Journal of Communication* 35, nr 3 (2020):213-229.

[9] Eric Goldman, „Content moderation remedies", *Michigan Technology Law Review* 28 (2021):1.

na lokalizację, stosowanie tymczasowych blokad, dodawanie etykiet sprawdzania faktów i ostrzeżeń śródtekstowych oraz pozbawianie użytkowników możliwości zarobkowania na platformach w ramach tzw. demonetyzacji[10]. Stosują je, argumentując zwykle, że dotyczą na przykład treści, które nie naruszają bezpośrednio wytycznych platformy, ale mogą podżegać do zachowań agresywnych.

Co istotne i podkreślane w artykułach, moderacja, w tym shadow baning, jest praktyką niezbędną i zważywszy na skalę naruszeń, polega na standardowych procedurach dalekich od rozpatrywania indywidualnych przypadków. Ludzcy moderatorzy, jeśli w ogóle są zaangażowani w ten proces, mają tylko chwilę na podjęcie decyzji i mogą kierować się pochopnymi osądami oraz ubogimi heurystykami. Nie mają czasu na ustalanie faktów ani na ważenie racji, dlatego niezbędna jest automatyczna moderacja[11]. Z kolei decyzje oparte na wnioskowaniu statystycznym mają niewiele lub nic wspólnego z ludzkim rozumowaniem wyrażonym w regułach opartych na języku[12].

2.2. ZŁOŻONOŚĆ SHADOW BANINGU

Jeśli spojrzy się na problem z perspektywy platformy i stosowanych przez nią rekomendacji, okazuje się, że proces jest skomplikowany także wówczas, gdy wykorzystywane są rozwiązania maszynowe. Po pierwsze, ponieważ problem dotyczy wielkich zasobów, identyfikacja „problematycznej" dla platformy treści odbywa się poprzez opracowanie klasyfikatora uczenia maszynowego, który może oszacować, jaką treść uznaje się za „problematyczną", oraz poprzez trenowanie tego klasyfikatora na wielu danych, które wcześniej zostały ocenione przez ludzkich moderatorów[13]. O praktyczno-technicznej stronie tego procesu wiadomo stosunkowo niewiele, ale na przykład Facebook korzysta z klasyfikatora uczenia maszynowego, który używał do identyfikowania treści „do usunięcia", YouTube opracował nowy klasyfikator treści, które uznaje za szkodliwe, oparty

[10] O zróżnicowanych formach monetyzacji na platformach, które stosują tzw. partnerskie programy dzielenia się dochodami (zwłaszcza YouTube, Twitch, Byte, DTube, Mixer), zob. Robyn Caplan i Tarleton Gillespie, „Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy". *Social Media+ Society*, nr 6: 1-13.

[11] Obszerny przegląd w: Vaishali U. Gongane, Mousami V. Munot i Alwin D. Anuse, „Detection and moderation of detrimental content on social media platforms: current status and future directions", *Social Network Analysis and Mining* 12, nr 1 (2022):129.

[12] Hannah Bloch-Wehba, „Automation in moderation", *Cornell Intitutional Law Journal* 53 (2020):41.

[13] Mary L. Gray i Siddharth Suri, *Ghost work: How to stop Silicon Valley from building a new global underclass* (New York: Harper Business, 2019).

na uczeniu maszynowym i szkolony przez tysiące ludzi[14]. Wiadomo także, że ludzcy moderatorzy korzystali z tych samych, dostępnych publicznie podręczników szkoleniowych (patrz: General Guidelines, 2025)[15], które Google używa do szkolenia oceniających jakość wyszukiwania. Po drugie, w ramach platform dochodzi do konfrontacji logik działania zespołów. Podczas gdy zespoły ds. zaufania i bezpieczeństwa („trust & safety") kierują się tym, co jest najmniej atrakcyjne dla odbiorców, zespoły zajmujące się systemami rekomendacji i kanałami informacyjnymi wybierają to, co jest uważane za najbardziej atrakcyjne.

Dodać należy, że proces degradacji nie zależy od jednego algorytmu, ale składa się z wielu fragmentarycznych jednostek obliczeniowych, które działają wspólnie, ale spełniają odrębne funkcje. Gdy jedne podsystemy optymalizują „zaangażowanie", inne optymalizują „zgodność". Ponadto całej tej aktywności towarzyszy przekonanie, że trudno określić stopień redukcji widoczności i nie wiadomo, do czego porównywać zmniejszoną widoczność treści. Nie sposób innymi słowy wyznaczyć „normalnego" poziomu zasięgu treści w sytuacji, gdy widoczność zależy od jej jakości, tego, kto ją zobaczył i polubił, jej popularności (oraz jej skali), tego, z czym treść tak konkurowała na platformie, jakie treści pojawiły się w tym samym czasie, jak oceniły je algorytmy uczenia maszynowego oraz jakie kryteria i progi były dostrojone w danym momencie[16].

2.3. NIEPRZEJRZYSTOŚĆ

Wspólnym mianownikiem artykułów pisanych na temat zarządzania platformami w zgodzie z krytyczną perspektywą badawczą jest to, że mechanizmy podejmowania decyzji są w dużej mierze ukrywane przed opinią publiczną praktycznie na każdym etapie moderacji, co skłania do ożywionego dyskursu obywatelskiego[17] i badań dotyczących nadawania znaczenia.

Przejrzystość dotyczy nie tylko tradycyjnych „przejrzystości aktora" (stron mogących wpływać na decyzje redakcyjne), źródła (stron, które dostarczają informację) i procesu redakcyjnego, w tym mechanizmów i decyzji redakcyjnych, ale także przejrzystości algorytmicznej. Ta ostatnia dotyczy zwykle danych

---

[14] Zoë Haime, *et al*., „Experiences of moderation, moderators, and moderating by online users who engage with self-harm and suicide content", *Digital Society* 4, nr 1 (2025):8.

[15] General Guidelines, accessed January 23, 2025 https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf.

[16] Daniel Trottier, *Social media as surveillance: Rethinking visibility in a converging world* (London: Routledge, 2016).

[17] Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media* (New Haven: Yale University Press, 2018).

wykorzystywanych w podejmowaniu decyzji, sposobu przetwarzania tych danych oraz wyników.

Wiedza użytkowników na tematy: jak, dlaczego lub nawet czy są w jakiś sposób związani z działaniem algorytmów, jest zazwyczaj ograniczona albo błędna. Nie wynika to jedynie z braku gotowości do jej pozyskiwania, ale także z ograniczonego dostępu do informacji. Tymczasem nieprzejrzystość algorytmiczna może ukrywać dyskryminację, umożliwiać manipulację lub powodować, że użytkownicy ślepo ufają podejmowaniu decyzji przez algorytmy[18].

W literaturze pokutuje pogląd, że o ile konwencjonalne metody usuwania i zawieszania kont są dobrze znane, a przede wszystkim zrozumiałe (co nie znaczy, że akceptowane), shadow baning należy z perspektywy użytkownika do największych wyzwań poznawczych, ponieważ zostawia mniej śladów albo nie zostawia ich wcale[19]. Usuwanie i zawieszenie konta jest dolegliwe, ale trudne do ukrycia przez platformę – może ona naturalnie skrywać swą decyzję i starać się utrzymywać użytkownika w przeświadczeniu, że jego treści są nadal widoczne, choć nikt, poza nim, ich nie widzi, ale takie rozwiązanie redukuje zaangażowanie innych użytkowników i wszyscy oni prędzej czy później zauważą problem, a ich podejrzenia będą łatwe do sprawdzenia. Shadow baning jest większym wyzwaniem.

3. SHADOW BANING NIE TYLKO JAKO JAKO CENZURA

Tak wielowymiarowa moderacja może budzić frustrację i bywa traktowana jako nieuczciwa[20], przyczyniając się do wzrostu nieufności użytkowników do całego procesu moderacyjnego oraz do poszczególnych jego narzędzi[21]. Oskarżenia dotyczą zwłaszcza demonetyzacji bez wyjaśnienia[22], uprzedzeń

[18] Brooke Erin Duffy i Colten Meisner, „Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility", *Media, Culture & Society* 45, nr 2 (2023):285-304.

[19] Mauro Conti, *et al*., „Revealing The Secret Power: How Algorithms Can Influence Content Visibility on Social Media", *arXiv preprint arXiv:2410.17390*, 2024.

[20] Sarah T. Roberts, „Digital detritus: 'Error' and the logic of opacity in social media content moderation", *First Monday* (2018).

[21] Samuel Mayworm, *et al*., „Content moderation folk theories and perceptions of platform spirit among marginalized social media users", *ACM Transactions on Social Computing* 7, nr 1-4 (2024):1-27.

[22] Emillie De Keulenaar, Anthony Glyn Burton i Ivan Kisjes, „Deplatforming, demotion and folk theories of Big Tech persecution", *Revista Fronteiras* 23, nr 2 (2021).

politycznych oraz marginalizacji określonych grup społecznych[23]. System moderacji nie jest neutralny politycznie i społecznie oraz rynkowo, a konstrukty, takie jak „zaangażowanie", „istotność" lub „jakość", mogą – jak sugeruje część badań – wydawać się obiektywne, ale w praktyce ich pomiar pociąga za sobą złożone i obciążone wartościami ważenie konkurujących interesów[24] i prowadzi do „władzy nad opinią publiczną"[25].

Redukcja widoczności staje bowiem nieustannie wobec znanych wyzwań i zarzutów: domniemanej władzy arbitra, a w istocie władzy elitarnych, prywatnych podmiotów nastawionych na zysk, możliwych uprzedzeń, niesprawiedliwego wpływu na różne społeczności użytkowników i implikacji dla wolności słowa[26], czy, w szerszym ujęciu, jako dowód cyfrowego autorytaryzmu („autorytaryzmu sieciowego", „techno-autorytaryzmu"), czyli wykorzystywania technologii do tłumienia praw i wolności[27].

Platformy ustanawiają zasady widoczności, ale same pozostają nieprzejrzyste. Ustalają warunki tego, co widoczne, ale same pozostają w cieniu i w efekcie dochodzić ma – jak zauważa Paddy Leerssen – do perwersyjnej odwrotnej proporcjonalności: najbardziej drastyczne przypadki naruszeń zasad formułowanych przez platformy są rozwiązywane za pomocą najmniej przejrzystych narzędzi: „zamiast Ministerstwa Prawdy dostajemy zatem, jak pisze Paddy Leerssen, Tajną Policję"[28].

Biorąc pod uwagę koncentrację władzy w gestii niewielkiej liczby firm mediów społecznościowych[29], ich zdolność regulowania wypowiedzi użytkowników jest nową i niepokojącą formą prywatnej „cenzury algorytmicznej", która może – zgodnie z foucaultowską koncepcją rządomyślności – umożliwić platformom społecznościowym sprawowanie bezprecedensowego stopnia kontroli nad

---

[23] Mark Werner, *et al*., „A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems", *International Journal of Human-Computer Studies* (2025):103468.

[24] Natali Helberger, „On the democratic role of news recommenders", *Algorithms, automation, and news*. (2021):14-33.

[25] Natali Helberger, *et al*., „Choice architectures in the digital economy: Towards a new understanding of digital vulnerability", *Journal of Consumer Policy* (2022):1-26.

[26] Daphne Keller, „Amplification and its discontents: Why regulating the reach of online content is hard", *Journal of Free Speech Law* 1 (2021):227.

[27] Roberts, Tony i Marjoke Oosterom, „Digital authoritarianism: a systematic literature review", *Information Technology for Development* (2024):1-25.

[28] Paddy Leerssen, „An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation", *Computer Law & Security Review* 48 (2023):105790.

[29] Jan Kreft, *Władza platform. Za fasadą Google, Facebooka i Spotify* (Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych Universitas, 2021).

komunikacją publiczną i prywatną, a uwolnienie z niej może być trudne bądź niepraktyczne[30]. Ponieważ ona jest również rutynowo wykorzystywana przez władze państwowe poprzez formalne zobowiązania prawne oraz różne rodzaje bardziej nieformalnej współpracy i wpływu, nie można wyraźnie rozróżnić cenzury publicznej od cenzury prywatnej[31].

3.1. AKCEPTACJA I OCZEKIWANIA PRZEJRZYSTOŚCI ORAZ PRZEJRZYSTEJ SZTUCZNEJ INTELIGENCJI

Z perspektywy użytkownika shadow baning budzi podejrzenia o nieuczciwą grę, dlatego niezbędne są starania o jego większą przejrzystość. Według części badaczy moderacja wydaje się dojrzałym krokiem platform i środowisko informacyjne może wymagać tego, co medioznawcy nazywają kuratelą. Należy jednak poznać, w jaki sposób – bez względu na ich motywację – platformy stosujące algorytmy rekomendacji oparte na sztucznej inteligencji wzmacniają jedne i redukują inne poglądy, jak przyczyniają się do chaosu w dyskusji obywatelskiej oraz jak może funkcjonować tzw. uczciwa sztuczna inteligencja[32].

Oczekiwana przejrzystość[33] moderacji przejawia się w publikowaniu raportów i bardziej szczegółowych publikacjach dotyczących zasad oraz oferuje mechanizmy odwołań od decyzji moderacji. Problem w tym, że techniki redukcji pozostają niewidoczne i dane na temat tego, co jest redukowane, nie są udostępniane publicznie. Poza nielicznymi wyjątkami nie są także określone zasady redukcji i jest szczególnie problematyczne, ponieważ redukcja nie pozostawia śladu.

3.2. ODSŁANIANIE TAJEMNICY

Choć w literaturze pojawiają się głosy, że shadow baning uniemożliwia jakiekolwiek akty oporu, w badaniach pojawiają się przynajmniej trzy drogi lepszego poznania jego mechanizmów.

---

[30] Jennifer Cobbe, „Algorithmic censorship by social platforms: Power and resistance", *Philosophy & Technology* 34, nr 4 (2021):739-766.

[31] Rachel Griffin, „The Politics of Algorithmic Censorship: Automated Moderation and its Regulation", *Music and the Politics of Censorship: From the Fascist Era to the Digital Age* (2025).

[32] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek i Sunil Aryal, „The pursuit of fairness in artificial intelligence models: A survey", *arXiv preprint arXiv:2403.17333* (2024).

[33] Max Z. Van Drunen, Natali Helberger i Mariella Bastian, „Know your algorithm: what media organizations need to explain to their users about news personalization", *International Data Privacy Law* 9, nr 4 (2019):220-235.

1) Ścieżka prawna, zwłaszcza w ramach Unii Europejskiej, wymuszająca na platformach większą przejrzystość.

2) Ścieżka pozyskiwania i wymiany wiedzy przez użytkowników, „mądrość tłumu" to luźne nawiązanie do słynnej pozycji Jamesa Surowieckiego[34]. To czerpanie z możliwości społeczności dzielącej się swymi doświadczeniami, domysłami, wyobrażeniami czy plotkami (algorytmicznymi), które niekiedy trafiają na rynek pod postacią recept samozwańczych „ekspertów" na shadow baning[35]. Jako tzw. teorie ludowe interpretowane także jako wyobrażenia algorytmiczne[36] i algorytmiczne plotki[37] skłaniają użytkowników do podejmowania „gier z algorytmami" lub do celowego, świadomego, indywidualnego „trenowania algorytmów".

3) Ścieżka trzecia to niszowe na razie rozwiązania techniczne w postaci tzw. zdecentralizowanych platform mediów społecznościowych (DSM) pozwalających użytkownikom stać się ich interesariuszami, w tym właścicielami swych kanałów[38].

3.3. SHDADOW BANING A PRAWO

Shadow baning jest obecny w stosunkowo bogatej literaturze naukowej dotyczącej regulacji prawnych. Uwaga ta dotyczy przede wszystkim DSA (Akt o Usługach Cyfrowych, ang. Digital Services Act) – pierwszych ważnych zapisów regulujących środki zaradcze dotyczące tzw. widoczności. Artykuł 1 DSA dotyczy moderowania treści wykraczających poza binarne rozwiązania „pozostaw – usuń". Akt ten wymaga, aby platformy skodyfikowały swoje zasady moderowania treści w „jasnym i jednoznacznym języku". Ponadto warunki korzystania z ich usługi muszą określać informacje o ograniczeniach, które nakładają na treści generowane przez użytkowników. Te informacje „powinny obejmować wszelkie polityki, procedury, środki i narzędzia używane w celu moderowania treści, w tym algorytmicznym podejmowaniu decyzji i dokonywanym przez

[34] James Surowiecki, *Mądrość tłumu: większość ma rację w ekonomii, biznesie i polityce* (Gliwice: Wydawnictwo Helion, 2010).

[35] Sophie Bishop, „Algorithmic experts: Selling algorithmic lore on YouTube", *Social Media+ Society* 6, nr 1 (2020):2056305119897323.

[36] Taina Bucher, „'Machines don't have instincts': Articulating the computational in journalism", *New Media & Society* 19, nr 6 (2017):918-933.

[37] Sophie Bishop, „Algorithmic experts: Selling algorithmic lore on YouTube", *Social Media+ Society* 6, nr 1 (2020):2056305119897323.

[38] Hamid Khobzi, Ana Isabel Canhoto i Mohammad Sadegh Ramezaniz, „Content creators at a crossroads between decentralized and centralized social media", *Business Horizons* 68, nr 1 (2025):109-120.

człowieka". Chodzi także o shadow baning, który jednak w DSA jest wymieniony tylko raz – art. 3(t)[39].

Reasumując, shadow baning pozostaje poważnym wyzwaniem normatywnym, ponieważ tajne sankcje, którymi w istocie jest, są trudniejsze do pociągnięcia do odpowiedzialności lub trudniej im się opierać. Z prawnej perspektywy nie można wobec nich występować. Tajne sankcje uniemożliwiają ich kwestionowanie, odwołanie, a zatem pozbawiają ofiary możliwości należytego procesu.

3.4. „GRY Z ALGORYTMAMI" I TEORIE LUDOWE

W obliczu niepewności co do intencji platform[40] oraz zagrożenia shadow baningiem użytkownicy podejmują starania, które górnolotnie nazwano „grą z algorytmami", „grą z platformą" bądź „trenowaniem algorytmu"[41]. Gry te podejmują zazwyczaj osoby, które wykonują twórczą pracę i pragną zachować dostęp na platformach, ale stały się ofiarami shadow baningu albo innych form „redukcji". W tym niezrównoważonym środowisku (na przykład na YouTube ok. 3% kanałów cieszy się oglądalnością na poziomie 90%) „gra z algorytmami" jest przedstawiana jako zjawisko wpisujące się w masowy wyzysk i kulturę przepracowania[42], a niepewność dochodów – jako pochodna widoczności wspieranej i ograniczanej przez systemy algorytmiczne[43].

---

[39] Zgodnie z DSA zasady przejrzystości, regulujące moderację treści przez pośredników internetowych, muszą być egzekwowane „w sposób staranny, obiektywny i proporcjonalny" oraz z należytym uwzględnieniem interesów i podstawowych praw. Co więcej, usługi pośredników muszą zawierać mechanizm powiadomień, za pomocą którego osoby trzecie mogą oznaczać treści do przeglądu w celu moderacji treści. Wreszcie, pośrednicy internetowi muszą dostarczać oświadczenie o powodach moderacji w przypadku każdej decyzji o moderacji treści.

[40] Sophie Bishop, „Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm", *Convergence* 24, nr 1 (2018):69-84.

[41] Ignacio Siles, *et al*., „Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South", *Big Data & Society* 7, nr 1 (2020):2053951720923377.

[42] Arturo Arriagada i Francisco Ibanez, „'You need at least one picture daily, if not, you're dead': Content creators and platform evolution in the social media ecology", *Social Media+ Society* 6, nr 3 (2020):2056305120944624; Zoe Zoe Glatt, „'We're all told not to put our eggs in one basket': uncertainty, precarity and cross-platform labor in the online video influencer industry", *International Journal of Communication* 16 (2022):3853-3871; Thomas Poell, David B. Nieborg i Brooke Erin Duffy, *Platforms and cultural production* (Oxford: John Wiley & Sons, 2021).

[43] Caitlin Petre, Brooke Erin Duffy i Emily Hund, „'Gaming the system': Platform paternalism and the politics of algorithmic visibility", *Social Media+ Society* 5, nr 4 (2019):2056305119879995; Victoria O'meara, „Weapons of the chic: Instagram influencer engagement pods as practices of resistance to Instagram platform labor", *Social Media+ Society* 5, nr 4 (2019):2056305119879671.

Użytkownicy starają się wówczas zachować tak zwaną suwerenność algorytmiczną[44]. Świadomi algorytmicznego wpływu nie chcą, aby o ich doświadczeniu decydowała sztuczna inteligencja i niechętnie kierują się decyzjami opartymi na algorytmach, a „przy okazji" kreują własne marki „ekspertów negocjacji". Zyskiwany niekiedy status autorytetu jest dodatkową motywacją negocjowania „suwerenności", balansowania między prywatnością i oczekiwaniem „zauważenia". „Ekspercka wiedza" jest wówczas sprzedawana w artykułach, filmowych instrukcjach oraz podczas warsztatów i jest konfrontowana z „tajemnicą algorytmiczną" oraz twierdzeniem platform, że nie ma żadnej tajemnicy.

Pozbywając się prywatności, użytkownicy uznają, że po części to nieunikniony koszt personalizacji doświadczeń i cena możliwości autoprezentacji. Dlatego dążą do poznania logiki algorytmu sztucznej inteligencji, analizują swoje doświadczenia i starają się zrozumieć, w jaki sposób mogą osiągnąć założone cele przez zmiany zachowań. „Trenując algorytm", starają się dostroić do jego wyobrażonych oczekiwań[45]. Niekiedy korzystają z narzędzi (np. Shadowban.eu i Whosban.eu) pozwalających w większym bądź w mniejszym stopniu automatycznie zidentyfikować shadow baning. Generalnie dążą do ustalenia sposobów na odzyskanie albo powiększenie widoczności swych treści i są przepojeni tzw. podejrzliwością algorytmiczną.

Gra z algorytmami podejmowana jest w oparciu o złożony, płynny i rizomatyczny kolaż tzw. teorii ludowych, wyobrażeń algorytmicznych i algorytmicznych plotek. Składają się one na krajobraz bardziej bądź mniej prawdopodobnych „przepisów na algorytmy" i mają dawać odpowiedź na pytanie, jak są zorganizowane kanały w mediach społecznościowych, jakie treści mają priorytet (albo są go pozbawione), jaki wpływ mają nowe treści oraz w jaki sposób łączą się z innymi użytkownikami. Formułując teorie ludowe i dzieląc się nimi, użytkownicy starają się poznać logikę maszyny i „normalizować" swoje zachowania pod kątem takich wyobrażonych zasad, a gdy te pozostają nadal nieprzeniknione, mogą skłonić się ku dystopijnym teoriom spiskowym i odrzucać argumenty o neutralności algorytmów, a losowe zdarzenia traktować jako dowody na istnienie nieprzychylnych wzorców[46].

---

[44] Urbano Reviglio i Claudio Agosti, „Thinking outside the black-box: The case for 'algorithmic sovereignty' in social media", *Social media+ society* 6, nr 2 (2020):2056305120915613.

[45] Hyunjin Kang i Chen Lou, „AI agency vs. human agency: understanding human – AI interactions on TikTok and their implications for user engagement", *Journal of Computer-Mediated Communication* 27, nr 5 (2022):014.

[46] Systematyczny przegląd literatury na ten temat np. w: Christine Bauer, *et al*., „Where are the values? A systematic literature review on news recommender systems", *ACM Transactions on Recommender Systems* 2, nr 3 (2024):1-40; Reza Jafari Ziarani i Reza Ravanmehr, „Serendipity in recommender systems: a systematic literature review", *Journal of Computer Science*

4. SHADOW BANING A ZARZĄDZANIE ALGORYTMICZNE

Zarządzanie algorytmiczne w środowisku mediów społecznościowych oznacza funkcjonowanie systemów społeczno-technicznych, które oceniają użytkowników i treści, przypisując im pozytywne znaczenie lub oceniając je jako (potencjalnie) ryzykowne lub „niedopuszczalne". W praktyce systemy platform służące do zarządzania treścią polegają na tym, że coraz częściej algorytmy automatycznie moderują albo same zgłaszają je moderatorom. Ta działalność polega nie tylko na rozpoznaniu tego, co „graniczne", ale także dostarczaniu poprzez tę selekcję algorytmom platform danych szkoleniowych, a tym samym uczenie się algorytmów, co na przykład jest klasyfikowane jako terroryzm czy mowa nienawiści.

W kontekście shadow baningu ten nurt badań jest marginalnie obecny, i to w niewielkiej puli artykułów naukowych. Część z nich odnosi się do szerszego pojęcia zarządzania platformami (poprzez groźby i kary) i konfiskaty tzw. kapitału społecznego gromadzonego przez użytkowników na platformach[47]. Inne dotyczą niewyjaśnialnych nawet dla projektantów systemów wyników zarządzania algorytmicznego. O ile bowiem ogólną logikę można wyjaśnić, poszczególne decyzje w ramach zarządzania algorytmicznego platformami, choćby wobec poszczególnych użytkowników, bywają dla nich niewytłumaczalne[48]. Jeszcze inne koncentrują się na zarządzaniu treściami w postaci zmian rekomendacji, rankingów lub wyników wyszukiwania[49].

Shadow baning pojawia się także w kontekście badań dotyczących tzw. zarządzania widocznością[50], które może przybierać różne formy i – jak zauważa Kokil Jaidka[51] – platformy mają cały arsenał środków, by usuwać treść z danej

---

and Technology 36 (2021):375-396; Jonathan Stray, *et al*., „Building human values into recommender systems: An interdisciplinary synthesis", *ACM Transactions on Recommender Systems* 2, nr 3 (2024):1-57; Jan Kreft, *Dziennik(AI)rstwo. Jak sztuczna inteligencja zmienia najciekawszą profesję* (Kraków: TAiWPN Universitas, 2025).

[47] Ori Schwartz, „Facebook rules: Structures of governance in digital capitalism and the control of generalized social capital", *Theory, Culture & Society* 36, nr 4 (2019):117-141.

[48] Kai Riemer i Sandra Peter, „Algorithmic audiencing: Why we need to rethink free speech on social media", *Journal of Information Technology* 36, nr 4 (2021):409-426.

[49] Blake Hallinan i Jed R. Brubaker, „Living with everyday evaluations on social media platforms", *International Journal of Communication* 15 (2021):40-96.

[50] Olga Kosińska, „Jej dokładność algorytmu: zarządzanie widzialnością a wykluczenie wizualne kobiet", w: *Ciało i umysł – konflikty, dialogi, reprezentacje*, red. Lidia Kamińska (Kraków: Wydawnictwo Naukowe Sub Lupa, 2023), 155-181.

[51] Kokil Jaidka, *et al*., „Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality", *Journal of Computer-Mediated Communication* 27, nr 1 (2022):zmab019.

funkcji („delisting"), zmniejszyć jej względną widoczność w ramach tej funkcji („demotion") lub ostrzegać przed treścią. Te ograniczenia mogą też być różne na różnych platformach, od „deslistingu" po „deboostowanie" (odpowiedzi na tweety innych użytkowników są niewidoczne). Możliwe jest również ograniczanie widoczności konkretnym użytkownikom, kohortom lub grupom demograficznym.

Tak wielopostaciowe zarządzanie widocznością jest także coraz bardziej złożonym procesem. Z perspektywy użytkownika wyzwanie jest tym większe, że widoczność na platformie stała się dostosowana do innych użytkowników. Gdy zatem treści są regulowane przez złożone systemy rekomendacji i wyszukiwania, shadow baning jest jednym z elementów i jako taki może być nie do identyfikacji[52]. Nic więc dziwnego, że shadow baning może być dostrzegany przede wszystkim wówczas, gdy dochodzi do gwałtownego spadku ruchu, choć i wtedy można ten skutek przypisać innym przyczynom, na przykład zwiększonej konkurencji o uwagę użytkowników.

PODSUMOWANIE

Z perspektywy kilku dekad wspomniany na wstępie rysunek z dwoma psami w roli głównej niewiele tłumaczy, co więcej przesłania istotę procesu zapośredniczonej komputerowo komunikacji i stosunków między jej podmiotami. Owszem, dla wielu użytkowników komunikacja z innymi członkami społeczności nadal przypomina bezpośrednią rozmowę „w realu", ale mit neutralności platform jako pośrednika uległ erozji[53].

Zastosowana w tym badaniu soczewka badawcza pozwoliła na uporządkowanie głównych nurtów dyskusji na temat shadow baningu.

Zidentyfikowana luka badawcza w postaci deficytu badań nad shadow baningiem z perspektywy zarządzania algorytmicznego stanowi obiecujący obszar dalszych badań. Uwzględniając tak ograniczony zakres badań, należy podkreślić że shadow baning to owoc polityki i architektury algorytmicznej, może zatem być interpretowany nie jako błąd w procesie walki z niepożądanymi praktykami użytkowników, ale jako element celowej strategii zarządzania platformami,

---

[52] Erwan Le Merrer, Benoît Mporgan i Gilles Tredan, „Setting the record straighter on shadow banning", *IEEE INFOCOM 2021-IEEE conference on computer communications* (Vancouver: IEEE, 2021).

[53] Brooke Erin Duffy i Colten Meisner, „Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility", *Media, Culture & Society* 45, nr 2 (2023):285-304.

mającej na celu kształtowanie społecznie i rynkowo motywowanego środowiska informacji, a pozyskiwane w tym środowisku wartościowe dane mogą dotyczyć przede wszystkim gier z algorytmami (jako konsekwencji wyobrażeń, teorii ludowych i plotek algorytmicznych). Taka interpretacja wydaje się uzupełniać interpretację moderacji ujętej w metaforach „ministerstwa prawdy" i „tajnej policji algorytmicznej" oraz utwierdzać w przekonaniu, że skupiają one uwagę i mogą skłaniać do pogłębionych badań[54].

Z zarządczej perspektywy enigmatyczne zasady dotyczące ryzykownych i niewłaściwych zachowań, umieszczane przez platformy w publicznych umowach i dokumentach, takich jak „warunki świadczenia usług", można interpretować jako celowo formułowane elastyczne ramy działania platform jako dostawców globalnych usług komercyjnych: środowiska kojarzenia reklamodawców i jednoczesnego tworzenia budowanych na zaufaniu relacji ze zróżnicowaną kulturowo i demograficznie bazą użytkowników. Shadow baning jest wówczas nie tylko remedium na zarzuty o stosowanie cenzury, ale narzędziem redukcji ryzyka utraty reputacji[55], także przez platformy, oraz unikania przez nie publicznej odpowiedzialności i pozyskiwania danych o tolerowanych przez użytkowników granicach ingerencji.

BIBLIOGRAFIA

Arriagada, Arturo i Francisco Ibanez. „You need at least one picture daily, if not, you're dead": Content creators and platform evolution in the social media ecology". *Social Media+ Society* 6, nr 3 (2020):2056305120944624.

Bauer, Christine, *et al*. „Where are the values? a systematic literature review on news recommender systems". *ACM Transactions on Recommender Systems* 2, nr 3 (2024):1-40.

Bishop, Sophie. „Algorithmic experts: Selling algorithmic lore on YouTube". *Social Media+ Society* 6, nr 1 (2020):2056305119897323.

Bishop, Sophie. „Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm". *Convergence* 24, nr 1 (2018):69-84.

Bishop, Sophie. „Algorithmic experts: Selling algorithmic lore on YouTube". *Social Media+ Society* 6, nr 1 (2020):2056305119897323.

Hallinan, Blake i Jed R. Brubaker. „Living with everyday evaluations on social media platforms". *International Journal of Communication* 15 (2021):19.

[54] Jan Kreft i Barbara Cyrek, „Kłamliwe, udane i błędne metafory sztucznej inteligencji chatbotów", *Roczniki Kulturoznawcze* 15 (2024):17-40.

[55] Kat Williams i Sebastian Bailey, „Online comment sections: Does taking them down enhance or hurt dialogue in a democracy?", *Journal of Media Ethics* 37, nr 4 (2022):285-287.

Bloch-Wehba, Hannah. „Automation in moderation". *Cornell Int'l LJ* 53 (2020):41.

Bucher, Taina. „'Machines don't have instincts': Articulating the computational in journalism". *New Media & Society* 19, nr 6 (2017):918-933.

Caplan, Robyn i Tarleton Gillespie. "Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy". *Social Media+ Society* 6 (2020):1-13.

Cobbe, Jennifer. „Algorithmic censorship by social platforms: Power and resistance". *Philosophy & Technology* 34, nr 4 (2021):739-766.

Conti, Mauro, *et al*. „Revealing The Secret Power: How Algorithms Can Influence Content Visibility on Social Media". *arXiv preprint arXiv:2410.17390*, 2024.

Cotter, Kelley. „Shadowbanning is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms". *Information, Communication & Society* 26, nr 6 (2023):1226-1243.

De Keulenaar, Emillie, Anthony Glyn Burton i Ivan Kisjes. „Deplatforming, demotion and folk theories of Big Tech persecution". *Revista Fronteiras* 23, nr 2 (2021): 118-139

Duffy, Brooke Erin i Colten Meisner. „Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility". *Media, Culture & Society* 45, nr 2 (2023):285-304.

Gillespie, Tarleton. „Do not recommend? Reduction as a form of content moderation". *Social Media+ Society* 8, nr 3 (2022):20563051221117552.

Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press, 2018.

Glatt, Zoe. „'We're all told not to put our eggs in one basket': uncertainty, precarity and cross-platform labor in the online video influencer industry". *International Journal of Communication* 16 (2022):3853-3871.

Goldman, Eric. „Content moderation remedies". *Michigan Technology Law Review* 28 (2021):1-60.

Gongane, Vaishali U., Mousami V. Munot i Alwin D. Anuse. „Detection and moderation of detrimental content on social media platforms: current status and future directions". *Social Network Analysis and Mining* 12, nr 1 (2022):129.

Gorwa, Robert, Reuben Binns i Christian Katzenbach. „Algorithmic content moderation: Technical and political challenges in the automation of platform governance". *Big Data & Society* 7, nr 1 (2020):2053951719897945.

Gray, Mary L. i Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. New York: Harper Business, 2019.

Griffin, Rachel. „The Politics of Algorithmic Censorship: Automated Moderation and its Regulation". *Music and the Politics of Censorship: From the Fascist Era to the Digital Age* (2025).

Grimmelmann, James. „The virtues of moderation". *Yale Journal Law & Technics* 17 (2015):42-109.

Haime, Zoë, *et al*. „Experiences of moderation, moderators, and moderating by online users who engage with self-harm and suicide content". *Digital Society* 4, nr 1 (2025):8.

Hallinan, Blake i Jed R. Brubaker. „Living with everyday evaluations on social media platforms". *International Journal of Communication* 15 (2021):1551-1569.

Helberger, Natali. „On the democratic role of news recommenders". *Algorithms, automation, and news* (2021):14-33.

Helberger, Natali, *et al*. „Choice architectures in the digital economy: Towards a new understanding of digital vulnerability". *Journal of Consumer Policy* (2022):1-26.

Jaidka, Kokil, *et al*. „Beyond anonymity: Network affordances, under deindividuation, improve social media discussion quality". *Journal of Computer-Mediated Communication* 27, nr 1 (2022):1-23.

Kang, Hyunjin i Chen Lou. „AI agency vs. human agency: understanding human – AI interactions on TikTok and their implications for user engagement". *Journal of Computer-Mediated Communication* 27, nr 5 (2022):014.

Keller, Daphne. „Amplification and its discontents: Why regulating the reach of online content is hard". *Journal of Free Speech Law* 1 (2021):227.

Kheya, Tahsin Alamgir, Mohamed Reda Bouadjenek i Sunil Aryal. „The pursuit of fairness in artificial intelligence models: A survey". *arXiv preprint arXiv:2403.17333* (2024).

Khobzi, Hamid, Ana Isabel Canhoto i Mohammad Sadegh Ramezaniz. „Content creators at a crossroads between decentralized and centralized social media". *Business Horizons* 68, nr 1 (2025):109-120.

Kosińska, Olga. „Jej dokładność algorytm: zarządzanie widzialnością a wykluczenie wizualne kobiet". W: *Ciało i umysł – konflikty, dialogi, reprezentacje*, red. Lidia Kamińska, 155-181. Kraków: Wydawnictwo Naukowe Sub Lupa, 2023.

Kreft, Jan. *Władza platform. Za fasadą Google, Facebooka i Spotify*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych Universitas, 2021.

Kreft, Jan. *Dziennik(AI)rstwo. Jak sztuczna inteligencja zmienia najciekawszą profesję*. Kraków: TAiWPN Universitas, 2025.

Kreft, Jan. *Władza platform. Za fasadą Google, Facebooka i Spotify*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych Universitas, 2021.

Kreft, Jan i Barbara Cyrek. „Kłamliwe, udane i błędne metafory sztucznej inteligencji chatbotów". *Roczniki Kulturoznawcze* 15 (2024):17-40.

Leerssen, Paddy. „An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation". *Computer Law & Security Review* 48 (2023):105790.

Mayworm, Samuel, *et al*. „Content moderation folk theories and perceptions of platform spirit among marginalized social media users". *ACM Transactions on Social Computing* 7, nr 1-4 (2024):1-27.

Le Merrer, Erwan, Benoît Mporgan i Gilles Tredan. „Setting the record straighter on shadow banning". *IEEE INFOCOM 2021-IEEE conference on computer communications*. Vancouver: IEEE, 2021: 1-10.

Nicholas, Gabriel. „Sunsetting 'Shadowbanning'". *Yale Law School Information Society Project Platform Governance Terminologies Essay Series* (2023):1-11.

O'Meara, Victoria. „Weapons of the chic: Instagram influencer engagement pods as practices of resistance to Instagram platform labor". *Social Media+ Society* 5, nr 4 (2019):2056305119879671.

Petre, Caitlin, Brooke Erin Duffy i Emily Hund. „'Gaming the system': Platform paternalism and the politics of algorithmic visibility". *Social Media+ Society* 5, nr 4 (2019):2056305119879995.

Petticrew, Mark i Helen Roberts. *Systematic reviews in the social sciences: A practical guide*. Oxford: John Wiley & Sons, 2008.

Poell, Thomas, David B. Nieborg i Brooke Erin Duffy. *Platforms and cultural production*. Oxford: John Wiley & Sons, 2021.

Reviglio, Urbano i Claudio Agosti. „Thinking outside the black-box: The case for 'algorithmic sovereignty' in social media". *Social media+ society* 6, nr 2 (2020):2056305120915613.

Riemer, Kai i Sandra Peter. „Algorithmic audiencing: Why we need to rethink free speech on social media". *Journal of Information Technology* 36, nr 4 (2021):409-426.

Roberts, Sarah T. „Digital detritus: 'Error' and the logic of opacity in social media content moderation". *First Monday* (2018).

Roberts, Tony i Marjoke Oosterom. „Digital authoritarianism: a systematic literature review". *Information Technology for Development* (2024):1-25.

Rogers, Richard. „Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media". *European Journal of Communication* 35, nr 3 (2020):213-229.

Savolainen, Laura. „The shadow banning controversy: perceived governance and algorithmic folklore". *Media, culture & society* 44, nr 6 (2022):1091-1109.

Schwartz, Ori. „Facebook rules: Structures of governance in digital capitalism and the control of generalized social capital". *Theory, Culture & Society* 36, nr 4 (2019):117-141.

Siles, Ignacio, *et al*. „Folk theories of algorithmic recommendations on Spotify: Enacting data assemblages in the global South". *Big Data & Society* 7, nr 1 (2020):2053951720923377.

Stockinger, Andrea, Svenja Schafer i Sophie Lechler. „Navigating the gray areas of content moderation: Professional moderators' perspectives on uncivil user comments and the role of (AI-based) technological tools". *New Media & Society* (2023):14614448231190901.

Stray, Jonathan, *et al*. „Building human values into recommender systems: An interdisciplinary synthesis". *ACM Transactions on Recommender Systems* 2, nr 3 (2024):1-57.

Surowiecki, James. *Mądrość tłumu: większość ma rację w ekonomii, biznesie i polityce*. Gliwice: Wydawnictwo Helion, 2010.

Trottier, Daniel. *Social media as surveillance: Rethinking visibility in a converging world*. London: Routledge, 2016.

Van Drunen, Max Z., Natali Helberger i Mariella Bastian. „Know your algorithm: what media organizations need to explain to their users about news personalization". *International Data Privacy Law* 9, nr 4 (2019):220-235.

Werner, Mark, *et al*. „A critical reflection on the use of toxicity detection algorithms in proactive content moderation systems". *International Journal of Human-Computer Studies* (2025):103468.

Willimas, Kat i Sebastian Bailey. „Online comment sections: Does taking them down enhance or hurt dialogue in a democracy?". *Journal of Media Ethics* 37, nr 4 (2022):285-287.

Ziarani, Reza Jafari i Reza Ravanmehr. „Serendipity in recommender systems: a systematic literature review". *Journal of Computer Science and Technology* 36 (2021):375-396.

SHADOW BANING: MIĘDZY „MINISTERSTWEM PRAWDY", „TAJNĄ POLICJĄ" ALGORYTMICZNĄ A POZYSKIWANIEM WIEDZY O UŻYTKOWNIKACH

Streszczenie

Gdy w 1993 roku *New Yorker* zamieścił rysunek z dwoma psami siedzącymi przed ekranem komputera z żartobliwym wpisem „W internecie nikt nie wie, że jesteś psem", każdy czytelnik wiedział, że choć chodzi anonimowość, dialog prowadzą ludzie. O roli cyfrowego pośrednika nie było ani słowa, a i nauka miała niewiele do zaoferowania na ten temat. Po ponad dwóch dekadach żart budzi życzliwy uśmiech, ale uwaga badaczy jest coraz częściej skupiona na „tym trzecim" – platformach cyfrowych, które dysponują złożoną wiedzą o użytkownikach, choć same występują w roli nieprzejrzystego co do intencji i możliwości, „ukrytego w ekranie" podmiotu ustalającego reguły zachowań i ekstrakcji wiedzy na platformie.

Przedmiotem przedstawianych badań jest shadow baning – rodzaj moderowania polegający na ukrywaniu/zmniejszaniu widoczności postów (użytkowników) przed wszystkimi użytkownikami oprócz ich autora. Jest to jedna z najbardziej opresyjnych form ingerencji platform mediów społecznościowych.

Przyjęta w badaniu metoda wspieranego maszynowo systematycznego przeglądu literatury pozwoliła na zidentyfikowanie nie tylko głównych nurtów badań nad shadow baningiem, ale rozpoznanie luki badawczej w postaci związku shadow baningu i zarządzaniem algorytmicznym. Shadow baning wyłania się w tej perspektywie nie tylko jako złożona, tajemnicza i opresyjna forma ingerencji platform, ale jako celowe działanie prowadzące do pozyskiwania wiedzy na temat granicy zaufania do platform.

Słowa kluczowe: shadow baning; moderacja; zarządzanie algorytmiczne; platformy mediów społecznościowych

SHADOW BANNING: BETWEEN THE "MINISTRY OF TRUTH", THE 'SECRET ALGORITHM POLICE' AND A STRATEGY FOR GETTING KNOWLEDGE ABOUT USERS

Summary

When, in 1993, the *New Yorker* posted a drawing of two dogs sitting in front of a computer screen with the humorous dialogue 'On the Internet, no one knows you're a dog,' every reader knew that, although anonymity was involved, people were having the dialogue. There was no word about the role of the digital intermediary, and science had little to offer on the subject either. More than two decades later, the joke raises a sympathetic smile, but the attention of researchers is increasingly focused on 'the third party' – digital platforms that hold complex knowledge about users, although they themselves act as a non-transparent entity as to intentions and capabilities, 'hidden in the screen' setting the rules of behaviour and knowledge extraction on the platform.

The subject of the research presented here is shadow banning – a type of moderation that involves hiding/reducing the visibility of posts (users) from all users except the author. This is one of the most oppressive forms of interference by social media platforms.

The method of machine-supported systematic literature review adopted in the study allowed the identification not only of mainstream research on shadow banning, but the recognition of a research gap in the form of the relationship between shadow banning and algorithmic governance. In this perspective, shadow banning emerges not only as a complex, mysterious and oppressive form of platform interference, but as a deliberate action leading to the acquisition of knowledge about the limit of trust in platforms.

Keywords: shadow banning; moderation; algorithmic management; social media platforms

ZAŁĄCZNIK 1

Are, Carolina. „An autoethnography of automated powerlessness: lacking platform affordances in Instagram and TikTok account deletions". *Media, Culture & Society* 45, nr 4 (2023):822-840.

Are, Carolina. „How Instagram's algorithm is censoring women and vulnerable users but helping online abusers". *Feminist Media Studies* 20, nr 5 (2020):741-744.

Are, Carolina i Pam Briggs. „The emotional and financial impact of de-platforming on creators at the margins". *Social media+ socjety* 9, nr 1 (2023):20563051231155103.

Conti, Mauro, *et al*. „Revealing The Secret Power: How Algorithms Can Influence Content Visibility on Social Media". arXiv preprint arXiv:2410.17390, 2024.

Cotter, Kelley. „Playing the visibility game: How digital influencers and algorithms negotiate influence on Instagram". *New media & socjety* 21, nr 4 (2019):895-913.

De Keulenaar, Emillie, Anthony Glyn Burton i Ivan Kisjes. „Deplatforming, demotion and folk theories of Big Tech persecution". *Revista Fronteiras* 23, nr 2 (2021).

Delmonaco, Daniel, *et al*. „'What are you doing, TikTok?'. How Marginalized Social Media Users Perceive, Theorize, and Prove Shadowbanning". *Proceedings of the ACM on Human-Computer Interaction* 8 (2024):1-39.

Duffy, Brooke Erin, *et al*. „The nested precarities of creative labor on social media". *Social media+ socjety* 7, nr 2 (2021):20563051211021368.

Duffy, Brooke Erin. „Algorithmic precarity in cultural work". *Communication and the Public* 5, nr 3-4 (2020):103-107.

Geboers, Marloes Annette i Chad Thomas Van De Wiele. „Regimes of visibility and the affective affordances of Twitter". *International Journal of Cultural Studies* 23, nr 5 (2020):745-765.

Gillespie, Tarleton. „Do not recommend? Reduction as a form of content moderation". *Social Media+ Society* 8, nr 3 (2022):20563051221117552.

Gorwa, Robert, Reuben Binns i Christian Katzenbach. „Algorithmic content moderation: Technical and political challenges in the automation of platform governance". *Big Data & Society* 7, nr 1 (2020):2053951719897945.

Horten, Monica. „Algorithms patrolling content: where's the harm?". *International Review of Law, Computers & Technology* 38, nr 1 (2024):43-65.

Jaidka, Kokil, Subhayan Mukerjee i Yphtach Lelkes. „Silenced on social media: the gatekeeping functions of shadowbans in the American Twitterverse". *Journal of Communication* 73, nr 2 (2023):163-178.

Johns, Amelia, *et al*. „Labelling, shadow bans and community resistance: did meta's strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread?". *Media International Australia* (2024):1329878X241236984.

Kuo, Tina, Alicia Hernani i Jens Grossklags. „The unsung heroes of facebook groups moderation: A case study of moderation practices and tools". *Proceedings of the ACM on Human-Computer Interaction* 7 (2023):1-38.

Leerssen, Paddy. „An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation". *Computer Law & Security Review* 48 (2023):105790.

Le Merrer, Erwan, Benoît Morgan i Gilles Trédan. „Setting the record straighter on shadow banning". W: *IEEE INFOCOM 2021-IEEE conference on computer communications*, 1-10. Vancouver: IEEE, 2021.

Monaci, Sara. „Media Technologies and Epistemologies: Platforming of Everything| The Governance of Disinformation: Everyday Practices of Platform Sovereignty". *International Journal of Communication* 18 (2024):16.

Moran, Rachel E., Izzi Grasso i Kolina Koltai. „Folk theories of avoiding content moderation: How vaccine-opposed influencers amplify vaccine opposition on Instagram". *Social Media+ Society* 8, nr 4 (2022):20563051221144252.

Marcondes, Francisco S., *et al*. „A profile on Twitter Shadowban: an AI ethics position paper on free-speech". W: *Intelligent Data Engineering and Automated Learning – IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings 22*, 397-405. Springer International Publishing, 2021.

Middlebrook, Callie. „The grey area: Instagram, shadowbanning, and the erasure of marginalized communities". Accessed October 15, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3539721.

Register, Yim, *et al*. „Attached to 'The algorithm': Making sense of algorithmic precarity on Instagram". W: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-15. New York: Association for Computing Machinery, 2023.

Risius, Marten i Kevin Marc Blasiak. „Shadowbanning: An Opaque Form of Content Moderation". *Business & Information Systems Engineering* (2024):1-13.

Savolainen, Laura. „The shadow banning controversy: perceived governance and algorithmic folklore". *Media, culture & society* 44, nr 6 (2022):1091-1109.

Stein, Jake, *et al*. „'You are you and the app. There's nobody else'. Building Worker-Designed Data Institutions within Platform Hegemony". W: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-26. New York: Association for Computing Machinery, 2023.

Théro, Héloïse i Emmanuel M. Vincent. „Investigating Facebook's interventions against accounts that repeatedly share misinformation". *Information Processing & Management* 59, nr 2 (2022):102804.