STUDIA PRAWA PUBLICZNEGO
2026 • NR 2 (54) ISSN 2300-3936
https://doi.org/10.14746/spp.2026.2.54.2

Martyna Kaczmarczyk*

Iluzja kontroli: czy *Human-in-the-Loop* spełnia wymogi realnego nadzoru człowieka w świetle prawa UE?

Illusion of Control: *Does Human-in-the-Loop* Meet the Requirements of Genuine Human Oversight under EU Law?

Abstract. The article provides a critical analysis of the Human-in-the-Loop (HITL) concept as a mechanism for ensuring human oversight over artificial intelligence systems under European Union law. The study aims to assess whether the presence of humans in the decision-making process genuinely fulfils the requirement of real or meaningful oversight, or whether it remains merely formal in character, giving rise to the so-called illusion of control. The analysis is based on both doctrinal and functional methods, including an interpretation of the provisions of the AI Act and the GDPR, particularly with regard to the obligation to ensure human oversight and the right to human intervention in automated decision-making processes.

The article argues that in many practical applications, HITL performs a largely symbolic function, often limited to the passive authorization of algorithmic decisions without any real capacity to question or modify them. This phenomenon is reinforced by factors such as automation bias, informational asymmetry, and the cognitive limitations of human operators. As a result, there is a significant risk of violations of individual rights, including the right to a fair procedure and the right to an effective remedy.

The article goes on to examine the legal implications of superficial oversight, particularly in the context of civil and administrative liability, and proposes criteria for assessing the effectiveness of HITL as a mechanism compliant with EU legal standards. It concludes by emphasizing the need to clarify the standard of "meaningful human control" and to introduce more stringent requirements regarding the actual role of humans in high-risk AI systems.

* University of Warmia and Mazury in Olsztyn, Poland | Uniwersytet Warmińsko-Mazurski w Olsztynie, Polska, https://orcid.org/0000-0001-6169-9466, e-mail: martyna.kaczmarczyk@uwm.edu.pl.

SPP 2(54), 2026: 27–43.

Keywords: Human-in-the-Loop – human oversight – artificial intelligence – legal liability – automated decision-making

Wprowadzenie

Postępująca integracja systemów sztucznej inteligencji (AI) z procesami decyzyjnymi prowadzi do istotnej transformacji sposobu funkcjonowania współczesnych instytucji publicznych i prywatnych. Algorytmy coraz częściej wspierają lub zastępują człowieka w zadaniach wymagających analizy danych, prognozowania czy kwalifikowania jednostek do określonych kategorii, co rodzi fundamentalne pytania o granice dopuszczalnej automatyzacji. W tym kontekście prawo Unii Europejskiej wyraźnie akcentuje konieczność zachowania kontroli człowieka nad działaniem systemów AI, traktując ją jako jeden z kluczowych instrumentów ochrony praw podstawowych. Wymóg ten znajduje odzwierciedlenie zarówno w regulacjach dotyczących systemów wysokiego ryzyka, takich jak AI Act[1], jak i w przepisach odnoszących się do zautomatyzowanego podejmowania decyzji, w szczególności w RODO[2].

Jednym z najczęściej wskazywanych modeli realizacji tego postulatu jest koncepcja *Human-in-the-Loop* (HITL), zakładająca włączenie człowieka w określone etapy funkcjonowania systemu algorytmicznego. Model ten bywa przedstawiany jako kompromis pomiędzy efektywnością automatyzacji a potrzebą zachowania kontroli i odpowiedzialności. Jednakże jego rzeczywista skuteczność pozostaje przedmiotem sporów. W praktyce bowiem udział człowieka może przyjmować bardzo różne formy – od aktywnej ingerencji w proces decyzyjny po czysto formalne zatwierdzanie wyników generowanych przez system.

---

[1] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2024/1689 z dnia 13 VI 2024 r. ustanawiające zharmonizowane przepisy dotyczące sztucznej inteligencji oraz zmiany rozporządzeń (WE) nr 300/2008, (UE) nr 167/2013, (UE) nr 168/2013, (UE) 2018/858, (UE) 2018/1139 i (UE) 2019/2144 oraz dyrektyw 2014/90/UE, (UE) 2016/797 i (UE) 2020/1828 (akt w sprawie sztucznej inteligencji) (Dz. Urz. UE L Nr 1689 z 12 VII 2024 r.), dalej „Akt w sprawie sztucznej inteligencji" lub „AI Act".

[2] Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 IV 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (Dz. Urz. UE L Nr 119 z z 4 V 2016 r., s. 1–88), dalej „Ogólne rozporządzenie o ochronie danych" lub „RODO".

Na tym tle zasadniczym problemem badawczym staje się ustalenie, czy obecność człowieka w takich konfiguracjach rzeczywiście zapewnia efektywny nadzór, czy też prowadzi do sytuacji, w której kontrola ma charakter jedynie deklaratywny. Wątpliwości te są szczególnie istotne w świetle rosnącej złożoności modeli AI oraz ograniczeń poznawczych i organizacyjnych po stronie użytkowników systemów. Artykuł podejmuje próbę uporządkowania tych zagadnień poprzez analizę normatywnych założeń nadzoru człowieka oraz ich konfrontację z realiami funkcjonowania systemów opartych na AI. Celem opracowania jest wykazanie, że bez precyzyjnego określenia standardów udziału człowieka istnieje ryzyko utrwalenia rozwiązań, które jedynie pozornie realizują wymogi prawa UE.

W zakresie metodologicznym artykuł opiera się przede wszystkim na metodzie dogmatycznoprawnej, polegającej na analizie i interpretacji obowiązujących przepisów prawa Unii Europejskiej, w szczególności wspomnianych AI Act oraz RODO. Analiza ta obejmuje zarówno wykładnię językową, jak i systemową oraz celowościową, co pozwala na odtworzenie normatywnego znaczenia pojęć takich jak „nadzór człowieka" (ang. *human oversight*) czy „znacząca kontrola" (ang. *meaningful human control*). Uzupełniająco zastosowano metodę prawnoporównawczą w ujęciu wewnątrzsystemowym, zestawiając regulacje dotyczące AI z istniejącymi mechanizmami ochrony praw jednostki w innych obszarach prawa UE.

Istotnym elementem badania jest również metoda funkcjonalna, umożliwiająca ocenę, w jakim stopniu przyjęte rozwiązania prawne odpowiadają rzeczywistemu sposobowi działania systemów sztucznej inteligencji. W tym kontekście wykorzystano ustalenia z zakresu nauk o zarządzaniu i informatyki, w szczególności dotyczące ograniczeń poznawczych użytkowników, zjawiska *automation bias* oraz praktyk implementacyjnych modelu *Human-in-the-Loop*. Pomocniczo odwołano się także do analizy przypadków, obejmujących wybrane zastosowania AI, takie jak *scoring* kredytowy, co pozwala na empiryczne zobrazowanie problemu pozorności nadzoru. Zastosowanie komplementarnego podejścia metodologicznego nie tylko umożliwia rekonstrukcję obowiązujących standardów prawnych, lecz także ich krytyczną ocenę pod kątem efektywności oraz adekwatności wobec dynamicznie rozwijających się technologii AI.

1. Koncepcja *Human-in-the-Loop* w ujęciu technicznym i prawnym

Koncepcja *Human-in-the-Loop* stanowi jedno z kluczowych zagadnień na styku prawa i technologii, odzwierciedlając próbę pogodzenia rosnącej automatyzacji procesów decyzyjnych z koniecznością zachowania kontroli człowieka nad systemami sztucznej inteligencji. Z perspektywy technicznej HITL odnosi się do takich architektur systemów AI, w których człowiek uczestniczy w określonym etapie ich działania. Udział ten może przyjmować różne formy – od oznaczania danych treningowych, przez walidację wyników, aż po podejmowanie ostatecznych decyzji na podstawie rekomendacji algorytmu[3]. W literaturze wyróżnia się także modele pokrewne, takie jak *Human-on-the-Loop*, gdzie człowiek monitoruje działanie systemu, oraz *Human-in-Command*, zakładający nadrzędną, strategiczną kontrolę człowieka nad jego funkcjonowaniem[4]. Kluczowe znaczenie ma przy tym nie tyle formalna obecność człowieka, ile zakres jego kompetencji oraz stopień faktycznego wpływu na wynik działania systemu.

Zastosowanie metody dogmatycznoprawnej pozwala zauważyć, że prawo Unii Europejskiej nie definiuje wprost pojęcia *Human-in-the-Loop*, lecz operuje kategoriami funkcjonalnie zbliżonymi, takimi jak „nadzór człowieka" (ang. *human oversight*)[5]. Pojęcie to odgrywa istotną rolę w regulacjach dotyczących sztucznej inteligencji, w szczególności w AI Act, gdzie stanowi jeden z warunków dopuszczalności stosowania systemów wysokiego ryzyka[6]. Regulacja ta wskazuje, że nadzór człowieka powinien umożliwiać zapobieganie ryzykom dla zdrowia, bezpieczeństwa oraz praw podstawowych lub ich minimalizowanie. Jednocześnie jednak przepisy nie precyzują w sposób jednoznaczny, jakie kryteria musi spełniać taki nadzór, aby można było uznać go za efektywny.

Podobne napięcia widoczne są na gruncie RODO, w którym art. 22 przyznaje jednostce prawo do niepodlegania decyzjom opartym wyłącznie na zautomatyzowanym przetwarzaniu oraz gwarantuje możliwość

[3] M.A. Goodrich, A.C. Schultz, *Human-Robot Interaction: A Survey*, „Foundations and Trends in Human–Computer Interaction" 2007, nr 1(3), s. 207–208.

[4] High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, Brussels 2019, s. 17–18.

[5] Szerzej: L. Edwards, *Regulating AI in Europe: four problems and four solutions*, London 2022, https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf (dostęp: 9 IV 2026).

[6] Akt w sprawie sztucznej inteligencji, art. 14.

uzyskania interwencji człowieka[7]. Wykładnia tych przepisów prowadzi do wniosku, że interwencja ta powinna mieć charakter rzeczywisty, a nie jedynie formalny. Jednak brak szczegółowych wytycznych powoduje, że w praktyce systemy mogą spełniać wymogi regulacyjne poprzez minimalne, często symboliczne zaangażowanie człowieka.

W świetle powyższych ustaleń należy stwierdzić, że koncepcja *Human-in-the-Loop*, choć normatywnie atrakcyjna, nie gwarantuje automatycznie realizacji celu, jakim jest zapewnienie realnej kontroli nad systemami AI. Jej skuteczność zależy od szeregu czynników, w tym od projektowania interfejsów, poziomu kompetencji użytkowników oraz przejrzystości działania systemu[8]. Brak uwzględnienia tych elementów może prowadzić do sytuacji, w której HITL pełni funkcję wyłącznie legitymizacyjną, maskując faktyczny brak kontroli człowieka. W konsekwencji zarówno analiza dogmatyczna, jak i funkcjonalna wskazują na potrzebę doprecyzowania standardów nadzoru człowieka w prawie UE. Niezbędne wydaje się wypracowanie kryteriów pozwalających odróżnić rzeczywisty nadzór od jego pozornej formy, co ma kluczowe znaczenie dla zapewnienia skutecznej ochrony praw jednostki w erze rosnącej automatyzacji.

2. Wymogi nadzoru człowieka w prawie UE

Wraz z rosnącym znaczeniem systemów sztucznej inteligencji w procesach decyzyjnych prawodawca unijny stanął przed koniecznością wypracowania mechanizmów zapewniających ochronę praw jednostki oraz ograniczenie ryzyk wynikających z automatyzacji. Jednym z centralnych instrumentów realizacji tego celu stał się wymóg nadzoru człowieka nad systemami AI. Analiza tego zagadnienia, przeprowadzona z wykorzystaniem metody dogmatycznoprawnej oraz funkcjonalnej, pozwala uchwycić zarówno normatywny zakres obowiązków nałożonych na podmioty wykorzystujące AI, jak i ich praktyczną skuteczność.

Z perspektywy dogmatycznoprawnej kluczowe znaczenie ma AI Act, który wprowadza zróżnicowane podejście do regulacji systemów AI w zależności od poziomu ryzyka. Regulacja wskazuje, że systemy AI wysokiego ryzyka powinny być zaprojektowane tak, aby człowiek mógł

[7] Ogólne rozporządzenie o ochronie danych, art. 22.

[8] T. Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, „Artificial Intelligence" 2017, nr 267(2), s. 7–8.

je faktycznie kontrolować i w razie potrzeby interweniować. Należy uniknąć sytuacji, w której człowiek formalnie „nadzoruje" system, ale w praktyce nie ma wpływu na jego działanie. Przepis wymienia różne formy nadzoru, np. możliwość monitorowania działania AI, zdolność do interpretowania wyników systemu, opcję zatrzymania lub zmiany działania systemu. Ważne jest też, żeby osoby sprawujące nadzór były odpowiednio przygotowane – powinny rozumieć, jak działa system i jakie niesie zagrożenia[9]. Jednocześnie jednak regulacja nie zawiera szczegółowych kryteriów pozwalających jednoznacznie ocenić, kiedy nadzór ma charakter efektywny, co pozostawia szerokie pole do interpretacji.

Uzupełnieniem powyższych regulacji są przepisy RODO, które – choć nie odnoszą się bezpośrednio do sztucznej inteligencji – mają istotne znaczenie dla oceny dopuszczalności zautomatyzowanego podejmowania decyzji. Zgodnie z art. 22 RODO osoba, której dane dotyczą, ma prawo do niepodlegania decyzjom opartym wyłącznie na zautomatyzowanym przetwarzaniu, jeżeli wywołują one wobec niej skutki prawne lub w podobny sposób istotnie na nią wpływają[10]. W takich przypadkach konieczne jest zapewnienie co najmniej prawa do uzyskania interwencji człowieka, wyrażenia własnego stanowiska oraz zakwestionowania decyzji. W doktrynie podkreśla się, że interwencja ta musi mieć charakter rzeczywisty, co oznacza, że osoba dokonująca oceny powinna być w stanie zrozumieć podstawy decyzji oraz podjąć autonomiczną decyzję, a nie jedynie potwierdzić wynik działania systemu[11].

Zastosowanie metody funkcjonalnej prowadzi jednak do wniosku, że realizacja tych wymogów w praktyce napotyka istotne trudności. Po pierwsze, współczesne systemy AI, zwłaszcza oparte na technikach

[9] Akt w sprawie sztucznej inteligencji, art. 14 ust. 4.

[10] Ogólne rozporządzenie o ochronie danych, art. 22 ust. 1. Takie przetwarzanie może być jednak dopuszczalne w określonych sytuacjach, np. gdy jest przewidziane prawem, niezbędne do wykonania umowy lub oparte na wyraźnej zgodzie osoby. W każdym przypadku musi być objęte odpowiednimi zabezpieczeniami, w tym prawem do interwencji człowieka, wyrażenia własnego stanowiska, uzyskania wyjaśnienia decyzji oraz jej zakwestionowania. Administratorzy danych są zobowiązani do stosowania rzetelnych metod przetwarzania, minimalizowania ryzyka błędów oraz zapobiegania dyskryminacji. Szczególnie rygorystyczne zasady dotyczą wykorzystywania w takich procesach danych wrażliwych, które mogą być przetwarzane tylko przy spełnieniu dodatkowych warunków.

[11] G. Malgieri, *Automated Decision-Making in the EU Member States: The Right to Explanation and Other 'Suitable Safeguards' in the National Legislations*, „Computer Law & Security Review" 2019, nr 5(35), artykuł 105327, s. 4–7.

uczenia maszynowego, charakteryzują się wysokim stopniem złożoności, co utrudnia ich interpretację przez użytkowników. W konsekwencji nawet formalnie zapewniony nadzór może być iluzoryczny, jeśli człowiek nie posiada narzędzi umożliwiających zrozumienie działania systemu. Po drugie, zjawisko *automation bias* powoduje, że użytkownicy mają tendencję do bezkrytycznego akceptowania rekomendacji algorytmów, co ogranicza ich skłonność do ingerencji. Po trzecie, czynniki organizacyjne, takie jak presja czasu, brak odpowiedniego przeszkolenia czy niewystarczające zasoby, dodatkowo osłabiają efektywność nadzoru[12].

Analiza funkcjonalna wskazuje również na problem asymetrii informacyjnej pomiędzy twórcami systemów AI a ich użytkownikami. Osoby sprawujące nadzór często nie mają dostępu do pełnych informacji o sposobie działania modelu, jego ograniczeniach czy o zakresie danych treningowych[13]. W takiej sytuacji trudno mówić o rzeczywistej kontroli, skoro decyzja człowieka opiera się na niepełnej wiedzy. Problem ten jest szczególnie istotny w kontekście realizacji praw jednostki, takich jak prawo do wyjaśnienia decyzji czy prawo do skutecznego środka odwoławczego.

Z perspektywy systemowej należy zauważyć, że wymóg nadzoru człowieka pełni w prawie UE funkcję gwarancyjną, mając przeciwdziałać dehumanizacji procesów decyzyjnych. Jednakże brak precyzyjnych standardów jego realizacji może prowadzić do sytuacji, w której spełnienie wymogów formalnych nie przekłada się na rzeczywistą ochronę jednostki. W tym kontekście szczególnego znaczenia nabiera potrzeba doprecyzowania pojęcia „znaczącej kontroli człowieka" (ang. *meaningful human control*), tak aby uwzględniało ono nie tylko formalną możliwość ingerencji, lecz także realne warunki jej wykonywania[14]. Analizy dogmatycznoprawna i funkcjonalna prowadzą do wniosku, że chociaż prawo Unii Europejskiej wyraźnie akcentuje znaczenie nadzoru człowieka nad systemami AI, to jego praktyczna skuteczność pozostaje ograniczona. Aby zapewnić realną ochronę praw jednostki, konieczne jest rozwinięcie bardziej precyzyjnych kryteriów oceny efektywności nadzoru oraz

[12] D.K. McGraw, *Ethical Responsibility in the Design of Artificial Intelligence (AI) Systems*, „International Journal of Responsibility" 2024, nr 7(1), artykuł 4, s. 11–12.

[13] B. Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, „Big Data & Society" 2016, nr 3(2), s. 2–3.

[14] F. Santoni de Sio, J. van den Hoven, *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, „Frontiers in Robotics and AI" 2018, nr 5(15), s. 1–3.

uwzględnienie czynników technologicznych i organizacyjnych, które determinują jego faktyczne funkcjonowanie.

3. Orzecznictwo dotyczące nadzoru człowieka nad systemami algorytmicznymi

Analiza orzecznictwa europejskiego i krajowego dotyczącego systemów algorytmicznych stanowi istotny punkt odniesienia dla oceny skuteczności koncepcji HITL jako mechanizmu zapewniającego nadzór człowieka. Chociaż nadal brakuje rozbudowanej linii orzeczniczej odnoszącej się bezpośrednio do tej koncepcji, w szczególności w kontekście AI Act, to jednak istnieją liczne rozstrzygnięcia pośrednio dotyczące problematyki kontroli nad decyzjami algorytmicznymi. Zastosowanie metod dogmatycznoprawnej oraz funkcjonalnej pozwala na identyfikację standardów wypracowywanych przez sądy oraz ocenę ich znaczenia dla interpretacji wymogu „realnego" nadzoru człowieka.

Z perspektywy dogmatycznoprawnej kluczowe znaczenie mają orzeczenia odnoszące się do ochrony praw podstawowych, w szczególności prawa do rzetelnej procedury oraz skutecznego środka odwoławczego. Zarówno Trybunał Sprawiedliwości Unii Europejskiej, jak i Europejski Trybunał Praw Człowieka podkreślają, że jednostka powinna mieć możliwość zakwestionowania decyzji, które wywołują wobec niej istotne skutki prawne. W kontekście systemów algorytmicznych oznacza to konieczność zapewnienia takiego modelu decyzyjnego, w którym człowiek nie pełni wyłącznie funkcji formalnej, lecz posiada realną zdolność do oceny i zmiany wyniku. Standard ten pozostaje w ścisłym związku z regulacjami RODO, w szczególności z art. 22, który gwarantuje prawo do interwencji człowieka w procesach zautomatyzowanego podejmowania decyzji[15].

Zastosowanie metody funkcjonalnej pozwala natomiast uchwycić, w jaki sposób sądy oceniają rzeczywiste działanie systemów algorytmicznych. Dobrym przykładem jest sprawa dotycząca systemu SyRI w Niderlandach, w której sąd uznał, że wykorzystanie narzędzi analitycznych do wykrywania nadużyć socjalnych narusza prawa jednostki

[15] Wyrok Trybunału Sprawiedliwości Unii Europejskiej z 7 XII 2023 r., sprawa C-634/21, *SCHUFA Holding AG*; wyrok Europejskiego Trybunału Praw Człowieka z 28 IV 2009 r., sprawa *K.H. i Inni przeciwko Słowacji*, skarga nr 32881/04; oraz Ogólne rozporządzenie o ochronie danych, art. 22.

ze względu na brak przejrzystości oraz niemożność skutecznego zakwestionowania decyzji[16]. Podobnie w sprawach dotyczących platform cyfrowych, takich jak Deliveroo, sądy krajowe wskazywały, że decyzje podejmowane na podstawie algorytmów mogą prowadzić do dyskryminacji, jeżeli brakuje efektywnych mechanizmów kontroli i weryfikacji[17].

W obu tych przypadkach kluczowym problemem nie była sama obecność człowieka w procesie decyzyjnym, lecz brak jego rzeczywistej sprawczości. Analiza funkcjonalna ujawnia, że formalne istnienie mechanizmu kontroli nie jest wystarczające, jeżeli człowiek nie dysponuje odpowiednimi narzędziami, wiedzą lub czasem na dokonanie rzetelnej oceny. Tym samym orzecznictwo pośrednio potwierdza tezę o ryzyku występowania „iluzji kontroli", w której nadzór człowieka ma charakter deklaratywny, a nie realny.

Warto również zwrócić uwagę na rozwijającą się linię orzeczniczą dotyczącą interpretacji art. 22 RODO. Sądy i organy nadzorcze podkreślają, że aby decyzja nie była uznana za „wyłącznie zautomatyzowaną", udział człowieka musi mieć charakter rzeczywisty i znaczący. Oznacza to, że osoba dokonująca oceny powinna być w stanie nie tylko formalnie zatwierdzić wynik, lecz także go zrozumieć i – w razie potrzeby – zmienić[18]. Taki kierunek interpretacji ma istotne znaczenie dla oceny modeli HITL, które w praktyce często ograniczają się do powierzchownej weryfikacji decyzji algorytmu.

Na tle powyższych ustaleń należy podkreślić, że brak bezpośredniego orzecznictwa dotyczącego AI Act nie oznacza braku standardów prawnych w tym zakresie. Przeciwnie, istniejące rozstrzygnięcia dostarczają istotnych wskazówek interpretacyjnych, które mogą być wykorzystane przy ocenie wymogu nadzoru człowieka przewidzianego w tej regulacji. W szczególności podkreślają one konieczność zapewnienia realnej, a nie jedynie formalnej kontroli nad działaniem systemów AI. Przegląd orzecznictwa prowadzi do wniosku, że europejskie standardy ochrony praw jednostki wymagają efektywnego nadzoru człowieka nad procesami decyzyjnymi, w tym również tymi wspomaganymi przez

[16] Wyrok Sądu Okręgowego w Hadze (Rechtbank Den Haag) z 5 II 2020 r., sprawa *NJCM i inni przeciwko Państwu Niderlandzkiemu (SyRI)*, ECLI:NL:RBDHA:2020:865.

[17] Wyrok Sądu w Bolonii (Tribunale di Bologna) z 31 XII 2020 r., sprawa *Rider Deliveroo (algorytm Frank)*, nr 2949/2020.

[18] Grupa Robocza Art. 29, *Wytyczne dotyczące zautomatyzowanego podejmowania decyzji i profilowania na potrzeby rozporządzenia (UE) 2016/679* (WP251rev.01, Bruksela, przyjęte 3 X 2017 r., zmienione 6 II 2018 r.), s. 21.

AI. Jednocześnie brak jednoznacznych kryteriów oceny tej efektywności stwarza ryzyko rozbieżności interpretacyjnych. W tym kontekście koncepcja *Human-in-the-Loop* powinna być oceniana nie przez pryzmat jej formalnej konstrukcji, lecz rzeczywistej zdolności do zapewnienia kontroli zgodnej z wymogami prawa UE.

4. Zastosowanie systemów sztucznej inteligencji

Analiza konkretnych zastosowań systemów sztucznej inteligencji stanowi istotne uzupełnienie rozważań teoretycznych dotyczących koncepcji *Human-in-the-Loop*. W niniejszym opracowaniu wykorzystano metody funkcjonalną oraz dogmatycznoprawną do zbadania systemu *scoringu* kredytowego, który stanowi jeden z najbardziej rozpowszechnionych przykładów algorytmicznego podejmowania decyzji o istotnym znaczeniu dla jednostki. Systemy te są szeroko stosowane przez instytucje finansowe do oceny zdolności kredytowej klientów, a ich decyzje mogą wywoływać daleko idące skutki prawne i ekonomiczne.

Z perspektywy technicznej system *scoringowy* opiera się na modelach uczenia maszynowego analizujących dane historyczne, takie jak historia kredytowa, dochody, struktura zobowiązań czy zachowania finansowe. Na tej podstawie generowany jest wynik (ang. *score*), który stanowi podstawę decyzji o przyznaniu lub odmowie kredytu. W wielu przypadkach systemy te są wspierane przez model HITL, w którym analityk kredytowy formalnie weryfikuje rekomendację algorytmu przed podjęciem ostatecznej decyzji[19].

Zastosowanie metody dogmatycznoprawnej pozwala ocenić ten model w świetle wymogów RODO, w szczególności art. 22, który odnosi się do zautomatyzowanego podejmowania decyzji. Kluczowe znaczenie ma tu pytanie, czy decyzja kredytowa rzeczywiście nie jest „wyłącznie zautomatyzowana", jeżeli człowiek uczestniczy w procesie jako podmiot zatwierdzający wynik. W doktrynie podkreśla się, że aby wyłączyć zastosowanie art. 22, interwencja człowieka musi mieć charakter rzeczywisty i znaczący[20]. Oznacza to, że analityk powinien posiadać kompetencje

[19] T. Khandani, A.J. Kim, A.W. Lo, *Consumer Credit-Risk Models via Machine-Learning Algorithms*, „Journal of Banking & Finance" 2010, nr 34(11), s. 2774–2775.

[20] S. Wachter, B. Mittelstadt, L. Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, „International Data Privacy Law" 2017, nr 7(2), s. 84–86.

oraz możliwość samodzielnej oceny sytuacji, a nie jedynie potwierdzać wynik wygenerowany przez system.

Analiza funkcjonalna ujawnia jednak, że w praktyce model HITL w systemach *scoringowych* często przyjmuje formę ograniczoną. Analitycy kredytowi działają w warunkach presji czasowej oraz organizacyjnej, co sprzyja automatycznemu akceptowaniu rekomendacji algorytmu. Dodatkowo złożoność modeli wykorzystywanych w *scoringu* – w szczególności modeli opartych na technikach uczenia maszynowego – utrudnia ich interpretację, co ogranicza zdolność do krytycznej weryfikacji wyników. W efekcie człowiek staje się jedynie ogniwem formalnym, a jego rola sprowadza się do zatwierdzania decyzji systemu[21]. Istotnym problemem jest również asymetria informacyjna. Instytucje finansowe często nie udostępniają pełnych informacji o sposobie działania modeli *scoringowych*, powołując się na tajemnicę przedsiębiorstwa. W rezultacie zarówno klienci, jak i sami analitycy mogą nie mieć dostępu do informacji niezbędnych do zrozumienia logiki decyzji[22]. Taka sytuacja rodzi poważne wątpliwości z punktu widzenia realizacji prawa do wyjaśnienia decyzji oraz skutecznego środka odwoławczego.

W świetle AI Act systemy *scoringu* kredytowego mogą być kwalifikowane jako systemy wysokiego ryzyka, co wiąże się z dodatkowymi obowiązkami w zakresie zapewnienia nadzoru człowieka. Regulacja ta wymaga, aby nadzór był zaprojektowany w sposób umożliwiający zapobieganie ryzykom dla praw jednostki, co implikuje konieczność zapewnienia realnej możliwości ingerencji w działanie systemu. Jednakże, podobnie jak w przypadku RODO, brak precyzyjnych kryteriów oceny efektywności nadzoru powoduje, że implementacja tych wymogów może mieć charakter formalny[23].

Przeprowadzona analiza wskazuje, że systemy *scoringu* kredytowego stanowią ilustrację szerszego problemu związanego z koncepcją *Human-in-the-Loop*. Chociaż formalnie zapewniają one udział człowieka w procesie decyzyjnym, to w praktyce jego rola bywa ograniczona przez czynniki technologiczne i organizacyjne. W konsekwencji istnieje ryzyko, że wymogi prawa UE dotyczące nadzoru człowieka będą realizowane w sposób pozorny, nie zapewniając rzeczywistej ochrony

[21] Z.C. Lipton, *The Mythos of Model Interpretability*, „Communications of the ACM" 2018, nr 61(10), s. 40–43.

[22] A.D. Selbst, J. Powles, *Meaningful Information and the Right to Explanation*, „International Data Privacy Law" 2017, nr 7(4), s. 233–241.

[23] Akt w sprawie sztucznej inteligencji, zał. III, art. 14.

praw jednostki. Skuteczność modelu HITL w systemach *scoringowych* zależy nie tylko od jego formalnej konstrukcji, lecz przede wszystkim od warunków jego implementacji. W celu zapewnienia zgodności z wymogami prawa UE konieczne jest doprecyzowanie standardów nadzoru oraz wprowadzenie mechanizmów zwiększających transparentność i realną sprawczość człowieka w procesie decyzyjnym.

5. Iluzja kontroli – analiza krytyczna i propozycje reinterpretacji

Koncepcja *Human-in-the-Loop*, choć stanowi jeden z centralnych mechanizmów regulacyjnych w obszarze sztucznej inteligencji, coraz częściej poddawana jest krytycznej analizie z perspektywy jej rzeczywistej efektywności. W wielu przypadkach nadzór człowieka nad systemami algorytmicznymi ma charakter pozorny, co prowadzi do powstania zjawiska określanego jako „iluzja kontroli". W szczególności problem ten ujawnia się w sytuacjach, gdy człowiek pełni funkcję tzw. *rubber stamp*[24], czyli podmiotu formalnie zatwierdzającego decyzje systemu bez ich rzeczywistej weryfikacji.

Z perspektywy dogmatycznoprawnej punkt wyjścia stanowią regulacje AI Act oraz RODO, które wprowadzają wymóg zapewnienia nadzoru człowieka nad systemami AI[25]. Analiza przepisów wskazuje, że prawodawca unijny przypisuje obecności człowieka funkcję gwarancyjną, mającą zabezpieczać prawa jednostki przed negatywnymi skutkami automatyzacji. Jednakże brak precyzyjnych kryteriów określających, kiedy nadzór można uznać za „znaczący", powoduje, że wymogi te mogą być realizowane w sposób minimalny. W praktyce oznacza to, że systemy mogą formalnie spełniać obowiązki regulacyjne poprzez włączenie człowieka w proces decyzyjny, nawet jeśli jego rola jest marginalna.

Zastosowanie metody funkcjonalnej pozwala pogłębić tę analizę poprzez odniesienie do rzeczywistych warunków funkcjonowania systemów AI. W wielu przypadkach udział człowieka ogranicza się do zatwierdzania decyzji wygenerowanych przez algorytm, bez realnej

[24] Oznacza podmiot formalnie decyzyjny, ale faktycznie pozbawiony realnej władzy. Szerzej: D. Keats Citron, *Technological Due Process*, „Washington University Law Review" 2008, nr 85(6), s. 1249–1313.

[25] Akt w sprawie sztucznej inteligencji, art. 14; Ogólne rozporządzenie o ochronie danych, art. 22.

możliwości ich zakwestionowania. Zjawisko to jest wzmacniane przez kilka czynników. Po pierwsze, złożoność modeli algorytmicznych utrudnia ich zrozumienie, co ogranicza zdolność do krytycznej oceny wyników. Po drugie, występuje tzw. *automation bias*, czyli tendencja do nadmiernego zaufania systemom technologicznym. Po trzecie, presja organizacyjna oraz ograniczenia czasowe powodują, że użytkownicy systemów nie są w stanie przeprowadzić pogłębionej analizy każdej decyzji[26]. W rezultacie nadzór człowieka staje się fikcją, a odpowiedzialność za decyzje pozostaje rozproszona. Konsekwencje tego zjawiska mają charakter zarówno prawny, jak i systemowy. Pozorny nadzór może prowadzić do naruszenia praw jednostki, w tym prawa do rzetelnej procedury oraz skutecznego środka odwoławczego. Ponadto utrudnia on przypisanie odpowiedzialności za błędne decyzje, gdyż formalna obecność człowieka może być wykorzystywana jako argument wyłączający odpowiedzialność za automatyczne przetwarzanie danych. W tym kontekście koncepcja HITL może paradoksalnie osłabiać, a nie wzmacniać ochronę jednostki.

W świetle powyższych ustaleń konieczne staje się podjęcie próby reinterpretacji obowiązujących standardów nadzoru człowieka. Z perspektywy dogmatycznej należy postulować odejście od formalnego rozumienia udziału człowieka na rzecz podejścia funkcjonalnego, w którym kluczowe znaczenie ma rzeczywista zdolność do ingerencji w proces decyzyjny. W tym kontekście pojęcie *meaningful human control* powinno być interpretowane jako obejmujące co najmniej trzy elementy: (1) kompetencje – człowiek musi posiadać wiedzę i umiejętności pozwalające na ocenę działania systemu; (2) informację – dostęp do danych i wyjaśnień umożliwiających zrozumienie decyzji; oraz (3) sprawczość – realną możliwość zmiany lub odrzucenia wyniku[27].

Zastosowanie metody funkcjonalnej prowadzi również do sformułowania konkretnych propozycji reform. Po pierwsze, konieczne jest wprowadzenie bardziej szczegółowych wymogów dotyczących projektowania systemów AI, które uwzględniałyby potrzeby użytkowników sprawujących nadzór. Obejmuje to m.in. rozwój narzędzi wyjaśnialnej sztucznej inteligencji (ang. *explainable AI*), które zwiększają przejrzystość działania modeli. Po drugie, istotne jest zapewnienie odpowiedniego poziomu szkolenia i wsparcia dla osób uczestniczących w procesie

[26] R. Parasuraman, D.H. Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, „Human Factors" 2010, nr 52(3), s. 387–389.

[27] F. Santoni de Sio, J. van den Hoven, op. cit., s. 2.

decyzyjnym. Po trzecie, należy rozważyć wprowadzenie mechanizmów selektywnego zaangażowania człowieka, polegających na jego aktywnym udziale w sytuacjach wysokiego ryzyka lub niepewności modelu, zamiast rutynowego zatwierdzania wszystkich decyzji[28].

Wreszcie, z perspektywy regulacyjnej zasadne wydaje się doprecyzowanie obowiązków podmiotów wykorzystujących AI poprzez wprowadzenie mierzalnych kryteriów oceny efektywności nadzoru. Mogłyby one obejmować m.in. częstotliwość ingerencji człowieka, jakość podejmowanych decyzji czy poziom zrozumienia działania systemu[29]. Takie podejście pozwoliłoby ograniczyć ryzyko nadużyć oraz zapewnić rzeczywistą realizację celów prawa UE. W tym kontekście istnieją standardy ISO/IEC, stanowiące międzynarodowe wytyczne dotyczące projektowania, wdrażania i oceny systemów informatycznych, w tym systemów opartych na sztucznej inteligencji. Standardy te są co do zasady dobrowolne. Nie mają mocy prawa same w sobie i organizacje nie muszą ich stosować. Szczególne znaczenie mają normy związane z zarządzaniem ryzykiem, bezpieczeństwem informacji oraz jakością procesów, które wspierają transparentność i odpowiedzialność działania algorytmów. Obowiązujące w standardach wskaźniki częstotliwości interwencji służą do monitorowania, jak często konieczna jest ingerencja człowieka w działanie systemu algorytmicznego. Mogą obejmować liczbę korekt decyzji, przypadków eskalacji do operatora lub sytuacji wymagających ręcznej weryfikacji[30]. Wysoka częstotliwość interwencji może wskazywać na ograniczoną skuteczność lub niezawodność algorytmu. Wprowadzenie wymienionych standardów jako obowiązujących norm prawnych wpłynęłoby na zwiększenie poziomu efektywności i rzeczywistego nadzoru człowieka nad systemami AI.

---

[28] E. Kamar, *Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence*, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI) 2016, s. 4071–4072, https://www.ijcai.org/Proceedings/16/Papers/603.pdf (dostęp: 1 V 2026).

[29] B. Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, „Nature Machine Intelligence" 2019, nr 1, s. 505–506.

[30] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), *ISO/IEC 42001:2023 Information Technology Artificial Intelligence Management System* (Geneva: ISO, 2023); International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), *ISO/IEC 23894:2023 Information Technology Artificial Intelligence Guidance on Risk Management* (Geneva: ISO, 2023); International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), *ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements* (Geneva: ISO, 2022).

Analizy dogmatycznoprawna i funkcjonalna wskazują, że koncepcja *Human-in-the-Loop*, w jej obecnym kształcie, nie zawsze gwarantuje efektywny nadzór człowieka nad systemami AI. Zjawisko *rubber stamp* stanowi poważne wyzwanie dla ochrony praw jednostki i wymaga zarówno reinterpretacji istniejących norm, jak i wprowadzenia nowych rozwiązań regulacyjnych. Tylko wówczas możliwe będzie zapewnienie, że obecność człowieka w procesie decyzyjnym ma charakter rzeczywisty, a nie jedynie iluzoryczny.

Podsumowanie

Przeprowadzona analiza koncepcji *Human-in-the-Loop* w świetle prawa Unii Europejskiej prowadzi do kilku zasadniczych wniosków o charakterze teoretycznym, normatywnym oraz praktycznym. Po pierwsze, należy stwierdzić, że prawo UE wyraźnie akcentuje znaczenie nadzoru człowieka jako instrumentu ochrony praw jednostki w warunkach rosnącej automatyzacji decyzji. Zarówno AI Act, jak i RODO przyznają obecności człowieka funkcję gwarancyjną, mającą przeciwdziałać arbitralności decyzji algorytmicznych. Jednakże analiza dogmatyczna wykazała, że regulacje te posługują się pojęciami o wysokim stopniu ogólności, takimi jak „nadzór człowieka" czy „interwencja człowieka", nie precyzując jednoznacznie kryteriów ich realizacji. W konsekwencji powstaje luka interpretacyjna, która umożliwia formalne spełnienie wymogów prawnych bez zapewnienia ich rzeczywistej skuteczności.

Po drugie, analiza funkcjonalna ujawniła, że w praktyce model HITL często nie spełnia przypisywanej mu roli. W wielu zastosowaniach człowiek uczestniczy w procesie decyzyjnym jedynie w sposób symboliczny, pełniąc funkcję *rubber stamp*, czyli podmiotu zatwierdzającego decyzje algorytmu bez ich pogłębionej weryfikacji. Zjawisko to jest wzmacniane przez czynniki takie jak złożoność modeli AI, asymetria informacyjna, ograniczenia poznawcze użytkowników oraz presja organizacyjna. W rezultacie dochodzi do powstania „iluzji kontroli", w której formalna obecność człowieka nie przekłada się na rzeczywistą zdolność do ingerencji w działanie systemu.

Po trzecie, analiza systemów *scoringu* kredytowego oraz przegląd orzecznictwa wskazują, że problem pozorności nadzoru ma charakter systemowy, a nie incydentalny. Sądy, choć nie odnoszą się bezpośrednio do koncepcji HITL, konsekwentnie podkreślają konieczność

zapewnienia realnej kontroli nad decyzjami wpływającymi na sytuację jednostki. Standard ten obejmuje nie tylko formalną możliwość interwencji, lecz także jej efektywność, rozumianą jako zdolność do zrozumienia i zakwestionowania decyzji. Tym samym orzecznictwo pośrednio wspiera tezę, że obecne modele HITL mogą być niewystarczające z punktu widzenia wymogów prawa UE.

Po czwarte, przeprowadzona analiza prowadzi do wniosku, że kluczowym problemem nie jest sama koncepcja *Human-in-the-Loop*, lecz sposób jej implementacji oraz brak precyzyjnych standardów oceny jej efektywności. W tym kontekście zasadne jest odejście od formalnego podejścia do nadzoru człowieka na rzecz podejścia funkcjonalnego, które uwzględnia rzeczywiste warunki działania systemów AI. Oznacza to konieczność redefinicji pojęcia *meaningful human control* poprzez wskazanie konkretnych kryteriów, takich jak poziom kompetencji użytkownika, dostęp do informacji oraz realna możliwość ingerencji w proces decyzyjny.

Po piąte, artykuł wskazuje na potrzebę wprowadzenia zmian o charakterze regulacyjnym i praktycznym. W szczególności postulowane jest doprecyzowanie obowiązków wynikających z AI Act poprzez określenie minimalnych standardów nadzoru człowieka oraz wprowadzenie mierzalnych wskaźników jego efektywności. Równocześnie konieczne jest rozwijanie narzędzi zwiększających przejrzystość systemów AI, takich jak rozwiązania z zakresu *explainable AI*, oraz zapewnienie odpowiedniego przygotowania osób odpowiedzialnych za nadzór. Istotnym kierunkiem rozwoju może być także model selektywnego angażowania człowieka, skoncentrowany na sytuacjach wysokiego ryzyka.

W konkluzji koncepcja *Human-in-the-Loop*, choć stanowi istotny element europejskiego modelu regulacji sztucznej inteligencji, w obecnym kształcie nie gwarantuje automatycznie realizacji celu, jakim jest zapewnienie realnego nadzoru człowieka. Bez doprecyzowania standardów prawnych oraz uwzględnienia uwarunkowań technologicznych i organizacyjnych istnieje ryzyko utrwalenia mechanizmów o charakterze iluzorycznym. W konsekwencji dalszy rozwój regulacji oraz praktyki stosowania prawa powinien zmierzać w kierunku wzmocnienia rzeczywistej roli człowieka w procesach decyzyjnych, tak aby odpowiadała ona nie tylko wymogom formalnym, lecz także funkcjonalnym.

BIBLIOGRAFIA

Edwards L., *Regulating AI in Europe: four problems and four solutions*, London 2022, https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf (dostęp: 9 IV 2026).

Goodrich M.A., Schultz A.C., *Human-Robot Interaction: A Survey*, „Foundations and Trends in Human–Computer Interaction" 2007, nr 1(3), s. 203–275.

Kamar E., *Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence*, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI) 2016, s. 4070–4073, https://www.ijcai.org/Proceedings/16/Papers/603.pdf (dostęp: 1 V 2026).

Keats Citron D., *Technological Due Process*, „Washington University Law Review" 2008, nr 85(6), s. 1249–1313.

Khandani T., Kim A.J., Lo A.W., *Consumer Credit-Risk Models via Machine-Learning Algorithms*, „Journal of Banking & Finance" 2010, nr 34(11), s. 2767–2787.

Lipton Z.C., *The Mythos of Model Interpretability*, „Communications of the ACM" 2018, nr 61(10), s. 36–43.

Malgieri G., *Automated Decision-Making in the EU Member States: The Right to Explanation and Other 'Suitable Safeguards' in the National Legislations*, „Computer Law & Security Review" 2019, nr 5(35), artykuł 105327, s. 1–26.

McGraw D.K., *Ethical Responsibility in the Design of Artificial Intelligence (AI) Systems*, „International Journal of Responsibility" 2024, nr 7(1), artykuł 4, s. 1–21.

Miller T., *Explanation in Artificial Intelligence: Insights from the Social Sciences*, „Artificial Intelligence" 2017, nr 267(2), s. 1–66.

Mittelstadt B., *Principles Alone Cannot Guarantee Ethical AI*, „Nature Machine Intelligence" 2019, nr 1, s. 501–507.

Mittelstadt B. et al., *The Ethics of Algorithms: Mapping the Debate*, „Big Data & Society" 2016, nr 3(2), s. 1–21.

Parasuraman R., Manzey D.H., *Complacency and Bias in Human Use of Automation: An Attentional Integration*, „Human Factors" 2010, nr 52(3), s. 381–410.

Santoni de Sio F., van den Hoven J., *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, „Frontiers in Robotics and AI" 2018, nr 5(15), s. 1–14.

Selbst A.D., Powles J., *Meaningful Information and the Right to Explanation*, „International Data Privacy Law" 2017, nr 7(4), s. 233–242.

Wachter S., Mittelstadt B., Floridi L., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, „International Data Privacy Law" 2017, nr 7(2), s. 76–99.