Nowoczesne Systemy Zarządzania
Zeszyt 18 (2023), nr 3 (lipiec-wrzesień)
ISSN 1896-9380, s. 45-84
DOI: 10.37055/nsz/183867

Instytut Organizacji i Zarządzania
Wydział Bezpieczeństwa, Logistyki i Zarządzania
Wojskowa Akademia Techniczna
w Warszawie

Modern Management Systems
Volume 18 (2023), No. 3 (July-September)
ISSN 1896-9380, pp. 45-84
DOI: 10.37055/nsz/183867

Institute of Organization and Management
Faculty of Security, Logistics and Management
Military University of Technology
in Warsaw

Propozycja wykorzystania uczenia przez wzmocnienie w celu optymalizowania podejmowania decyzji w zakresie przeciwdziałania praniu pieniędzy oraz finansowania terroryzmu (część 1)

A proposal to use reinforcement learning to optimize decision-making in the field of counteracting money laundering and terrorist financing (Part 1)

Maciej Aleksander Kędzierski
sulawezi.mk@onet.eu; ORCID: 0000-0003-3074-1355

Abstrakt. Uczenie przez wzmocnienie stanowi propozycję do rozwiązywania problemów identyfikacji i weryfikacji klientów instytucji obowiązanych, którzy mogą być powiązani z procederem prania pieniędzy czy finansowaniem terroryzmu. Może to mieć zastosowanie zarówno na poziomie czynności weryfikacyjnych, jak i na poziomie monitoringu klienta danej instytucji. Model uczenia przez wzmocnienie pozwala na uzyskiwanie rezultatów akcji agenta jako nie tylko konsekwencji jego uczenia, lecz także podejmowania własnych decyzji zmierzających do uzyskania jak największej nagrody. Wsparciem tego typu działań jest dostarczanie danych technicznych, a także współpraca z czynnikiem ludzkim w ramach uczenia się ze wzmocnieniem na podstawie informacji zwrotnej od ludzi. Oprócz samej idei włączenia tego typu modelu myślenia maszynowego na poziom analityki instytucji obowiązanej pozostaje także uzyskiwanie za jego pośrednictwem rezultatów w postaci predykcyjnego wykrywania zagrożenia związanego z możliwością legalizowania środków przestępczych i inwestowania ich w działalność terrorystyczną.
Słowa kluczowe: uczenie przez wzmacnianie, pranie pieniędzy, model Markowa, agent, zbiór uczący, sprzężenie zwrotne

Abstrakt. Reinforcement learning is a proposal for solving the problems of identifying and verifying customers of mandatory institutions who may be connected with money laundering or terrorist financing. Its application can take place both at the level of verification activities but also at the level of monitoring of the institution's client. The reinforcement learning model allows the results of an agent's actions to be obtained as not only a consequence of his learning, but also of his own decision-making aimed at obtaining the greatest possible reward. Supporting this type of action is not only the provision of technical data but also the collaboration with the human agent in Reinforcement Learning from Human Feedback. In addition to the very idea of incorporating this type of machine thinking model into the analytical level of

the obligated institution, it remains to obtain results through it in the form of predictive threat detection related to the possibility of legalizing criminal funds and investing them in terrorist activities.
Keywords: reinforcement learning, money laundering, Markov model, agent, training set, feedback

Wprowadzenie

Dotychczasowe rozwiązania w zakresie wykorzystania sztucznej inteligencji (ang. *artificial intelligence*, AI) i uczenia maszynowego (ang. *machine learning*, ML) stały się z czasem kanonem wykorzystywanym przez instytucje obowiązane w obszarze przeciwdziałania praniu pieniędzy oraz finansowaniu terroryzmu AML/CFT (ang. *Anti-Money Laundering/Counter Financing of Terrorism* – przeciwdziałanie praniu pieniędzy oraz finansowaniu terroryzmu). To, wydające się już być trywialnym stwierdzenie, między innymi jest konsekwencją tego, że stosowane dotychczas predefiniowane konwencjonalne reguły czy schematy przeciwdziałania były łatwo rozpoznawalne, a tym samym osoby zajmujące się ML/FT mogły łatwo dostosowywać swoje zachowanie, tak aby uniknąć wykrycia czynu, a w konsekwencji prowadzić, w dłuższym okresie, przestępczy proceder. System przeciwdziałania funkcjonował w taki sposób, że opracowywano uprzednio określone reguły, które konfrontowano z danymi wejściowymi, i obserwowano, czy nie występują w nich te ustalone reguły-wzorce. W konsekwencji, po takim przetworzeniu danych, wygenerowywano dane wyjściowe (wynik), identyfikując zdarzenie z użyciem przyjętych cech. Podejście to powodowało znaczną liczbę wyników jako schemat zgodny z regułami, jednakże bez wartości dla wykrycia podejrzanego procederu ML/FT (wyniki fałszywie pozytywne). Z czasem zmieniono podejście, opierając je na zasadzie ryzyka, metodach przeciwdziałania, które wymagały, zwłaszcza w rozbudowanych produktowo (usługowo) i organizacyjnie instytucjach, stałego przetwarzania znacznej liczby danych (metadanych) i prowadzenia dynamicznej aktualizacji wyników. W przypadku nadzorowanego ML sytuacje są znane jako wejścia i wyjścia, co ewentualnie umożliwia zastosowanie wspomnianego ML do powielanych schematów postępowania dla krótkich łańcuchów powiązań zdarzeń oraz ustalonego efektu końcowego. Przykładowo mogą go stosować instytucje obowiązane (IO) świadczące usługi/produkty spełniające te warunki (np. notariusze). Dla bardziej skomplikowanych produktów/usług oraz wielości konfiguracji świadczonych przez IO na skalę powodującą generowanie znacznej liczby danych w krótkiej jednostce czasowej bądź długie łańcuchy przestępcze niezbędne staje się skorzystanie z innych możliwości, które daje uczenie maszynowe, np. w postaci nienadzorowanego uczenia maszynowego (zwłaszcza dotyczy to instytucji bankowych).

Do wskazanych możliwości ML dołącza kolejna metoda, a w zasadzie rozwiązanie pośrednie, to jest uczenie przez wzmacnianie (ang. *reinforcement learning*, RL). Uczenie przez wzmocnienie stanowi podzbiór uczenia maszynowego (ML), w którym algorytm jest zasilany nieoznaczonym zestawem danych, aktywny agent

wybiera działanie dla każdego punktu danych i otrzymuje informację zwrotną, co pomaga agentowi uczyć się i wybierać najlepszą opcję. RL różni się od uczenia nadzorowanego (będącego podzbiorem ML) tym, że w przypadku uczenia nadzorowanego dane szkoleniowe zawierają „klucz odpowiedzi", a model trenowany jest z samą poprawną odpowiedzią. W przypadku RL nie ma odpowiedzi, natomiast agent wzmacniający decyduje, co zrobić, aby jak najlepiej wykonać zadanie. Stąd też do realizacji tego zadania wykorzystywane jest sprzężenie zwrotne, w którym uczestniczy agent oraz funkcjonuje metoda „nagradzania" (nagroda dodatnia) lub „karania" (nagroda ujemna) i działanie w scenariuszu „prób i błędów" (celem którego staje się maksymalizacja nagrody). Agent, idąc w dobrym kierunku, uzyskuje nagrodę, oddalając od siebie karę. Agent kontynuujący dobry kierunek uzyskuje coraz większą nagrodę ($G_t = R_{t+1} + R_{t+2} + \ldots$). Ponadto RL jako element ML mieści się w pojęciu analizy zaawansowanej (ang. *Advanced Analysis*, AA). RL wykorzystuje tylko otoczenie (środowisko), z którego dane są zbierane automatycznie do bufora. Tym samym dla RL będzie ustalone środowisko (ang. *environment*), z którego model będzie zbierał dane automatycznie. Funkcja, która opisuje akcję, jest określana jako polityka. Agent ma na celu znalezienie optymalnej polityki, którą maksymalizuje nagroda kumulacyjna (zdyskontowana), co oznacza, że nagroda może być przyznawana na przykład za wykrywanie alertów (wykrycia zagrożenia). Szczególnie ważne jest wprowadzenie RL do procesów AML, jako uczenia nienadzorowanego, wobec braku wystarczających danych na potrzeby akcji agenta. Niemniej jednak występują też dwa krytyczne punkty RL, a jest to wybór polityki i aproksymacja funkcji oznaczająca najściślejsze dopasowanie funkcji w danej klasie do funkcji docelowej w sposób specyficzny odpowiadającej danemu zadaniu.

AI jest obecnie typowana jako instrument do negowania zasady proporcjonalności wykorzystywania danych w związku z przeciwdziałaniem ML/FT. Niemniej jednak istotne jest, aby uzyskać właściwą równowagę między walką z przestępczością a prywatnością. Ta de-równowaga widoczna jest przede wszystkim w ilości danych przekazywanych do jednostek analityki finansowej (JAF) i efektywności wykorzystania tych danych na rzecz rzeczywistego wykrycia procederu prania pieniędzy czy finansowania terroryzmu. W tym kontekście nie wydaje się, że AI czy RL byłoby bardziej groźne wobec nieprzestrzegania zasady proporcjonalności niż nieprzygotowany pracownik IO podejmujący decyzje, w wyniku których następuje niewspółmierny do ryzyka i zagrożenia eksport danych do JAF, zwłaszcza że AI z wykorzystaniem zbiorów uczących jest jedynie wprowadzeniem pomocniczego instrumentu (wydaje się, że jest bardziej wydajny od personalnego) przestrzegania reżimu określonego w IO modelowego rozwiązania dla świadczonych produktów i usług. W konsekwencji wykorzystania AI umożliwia ograniczanie zdarzeń przypadkowych (bezużytecznych) kwalifikowanych jako podejrzane wobec potrzeby przekazania o nich informacji do JAF.

Należy też pamiętać, że w związku z obecnymi rozwiązaniami, o czym wspomniano na wstępie, zwłaszcza zawartymi w ustawie z dnia 1 marca 2018 r. o przeciwdziałaniu praniu pieniędzy oraz finansowaniu terroryzmu (dalej jako: ustawa o p.p.p.f.t.), ostatecznym decydentem nie jest „maszyna", lecz „człowiek". RL staje się ważnym elementem wsparcia optymalizacji procesu decyzyjnego decydenta/nadzorcy – AMLO/MLRO (ang. *Anti-Money Laundering Officer/Money Laundering Reporting Officer* – oficer przeciwdziałający praniu pieniędzy/oficer raportujący o zdarzeniach jako powiązanych z praniem pieniędzy) w IO na rzecz wysłania jedynie wyeksponowanego racjonalnie raportu SAR do JAF. Przede wszystkim dotyczy to etapu pracy wykonanego przed podjęciem decyzji o przekazaniu informacji poza układ, jakim jest sama IO. Zastosowanie RL ma miejsce – jako reakcja – gdy działanie/sekwencję działań w dowolnym środowisku oparto na metodzie prób i błędów. Takie podejście jest więc klasycznym odejściem jedynie od odchyleń „modelowych" na rzecz inwencji sprawców przestępstwa, którzy sami kreują nowe zdarzenia niezidentyfikowane przez analityków IO. Można powiedzieć, że RL jest techniczną odpowiedzią na „intelektualną" inwencję twórczą sprawców przestępstw. Zwłaszcza tych sprawców, których taktyka przestępcza nie jest oparta na powielaniu schematów kryminalnego działania celem uniemożliwienia organom ścigania wykrycia sprawcy za pośrednictwem technik kryminalistycznych i analizy kryminalnej, głównie w zakresie pozostawiania w podobnym schemacie śladów kryminalistycznych lub śladów w przestrzeni teleinformatycznej, posługując się na przykład mobilną bankowością. Zastosowanie RL oznacza postęp pomiędzy uczeniem się na „własnych ocenach" a uczeniem się na „własnych błędach". Ponadto należy być przygotowanym – jednak na zasadzie ogólności – gdy sprawcą okazuje się osoba „ułomna", wytypowana przez organizatorów przestępstwa ML/FT właśnie ze względu na swoją ułomność – wtedy wobec braku możliwości dotarcia do środowiska i korelowanego zagrożenia należałoby zastosować metodę RL jako „jedynie adekwatną" w stosunku do stwierdzonych działań osoby fizycznej. Niestety należy być zawsze przygotowanym na „innowacyjność" działania sprawców, którzy zechcą przejąć lub wykorzystać IO na potrzeby dokonania procederu ML/FT. Dlatego też każda metoda sięgająca głęboko do zachowania i wiedzy „przestępczego decydenta" w konfiguracji z możliwościami IO staje się bezcenna wobec potrzeby przeciwdziałania. Stąd zastosowanie RL na etapie rozpoznawania zagrożenia przed sformułowaniem i skierowaniem SAR (ang. *Suspicious Activity Report* – raport o podejrzanej aktywności) do JAF. Zdecydowanie dzisiejsze bazowanie na odchyleniach od przyjętych (przez IO) modeli postępowania z produktem lub ofiarowaną usługą nie jest właściwe wobec inwencji sprawców. W konsekwencji należałoby zaproponować zastosowanie innej metody, np. metody RL.

Powodów do konstrukcji wsparcia procesu decyzyjnego opartych na RL w systemach AML/CFT jest kilka. Można wymienić następujące:

– zmienność w czasie rodzajów czynników i natężenia czynników ryzyka indywidualnego w obszarze ML/FT;

- wysoki poziom uznawania za atut efektywności zagrożenia ryzyka zapytań kierowanych przez JAF i jednostki współpracujące do IO bez możliwości weryfikacji rzeczywistych potrzeb tych zapytań;
- występujące niekiedy (zwłaszcza w momencie wdrażania nowych produktów) trudności w kwestii możliwości sprowadzania stanu niepewności do poziomu prawdopodobieństwa ryzyka;
- potrzeba realnej oceny niemożliwości zastosowania określonych ustawowo środków bezpieczeństwa finansowego w relacji z klientem, czego konsekwencją może być nienawiązanie relacji lub ich rozwiązanie;
- nieskuteczność kontroli i audytu w IO wobec niskiej efektywności realizacji i wdrażania systemu AML/CFT w danej instytucji;
- brak ciągłości wiedzy, pamięci instytucjonalnej oraz wysoki poziom rotacji pracowników komórek AML/CFT w IO;
- bazowanie w ocenie analitycznej IO jedynie na wypracowanych lub zadysponowanych schematach weryfikacji taktyki działań sprawców przestępstw (schematyzacji analizy);
- niewłaściwe przygotowanie wzorca prawidłowego postępowania wobec nowych produktów, a tym samym błędne ujawnianie anomalii;
- potrzeba wzmocnienia technicznego już funkcjonujących projektów w IO w zakresie rozpoznawania ML/FT nieopartych na RL;
- wysoka częstotliwość wdrażania w bankach (jako głównych IO dostarczających danych z zakresu AML/CFT do JAF) systemów opartych na AI w obszarze obsługi klienta;
- potrzeba przygotowania procesu decyzyjnego AML/CFT do przestrzennej organizacji finansów (przepływów środków).

Należy także zauważyć, że skorzystanie z RL wymagać będzie od IO zachowania pewnej stałej dyscypliny co do ustalania i grupowania danych o podejrzanym działaniu (naznaczania elementów środowiska). Ma to związek z potrzebą identyfikowania modeli schematów przestępczych transakcji i cechowania (również tymi transakcjami) sprawców – uczestników tych schematów. Ponadto należy uznać, że sam aktywny „związek" przestępczy (jako porozumienie w celu popełniania przestępstw), mimo potrzeby monitoringu wielu transakcji, jest znacznie ograniczonym personalnie, pod względem inicjacyjnym i wykonawczym, podmiotem gospodarczym, a także przynależnymi mu instrumentami finansowymi, w tym kontami, którymi chce posłużyć się sprawca przestępstwa. Pozwala to w sposób szczególny identyfikować działalność niezgodną z prawem. Wyłomem mogą być tu jedynie przestępstwa podatkowe (wyłudzenia podatku VAT), budowane kompleksowo również na bazie podmiotów trzecich, które są obserwowalne przez IO na kontach, a dodatkowo za pomocą systemu STIR (System Teleinformatyczny Izby Rozliczeniowej).

Podstawowy model zastosowania metody RL na potrzeby AML/CFT będzie składał się z kolejnych kroków: określenie środowiska, zdefiniowanie nagrody, wyznaczenie agenta, trenowanie agenta oraz wdrożenie rozwiązania i jego doskonalenie.

W przypadku takiego podejścia przed praktycznym zastosowaniem RL niezbędne jest zdefiniowanie kilku zagadnień:

- opisanie samego środowiska, w którym będzie działał agent (algorytm, interfejs agent – środowisko);
- określenie nagrody jako ustalonego impulsu pozytywnie rozpoznawalnego przez agenta w przypadku „hitu" (polega na dostarczaniu pozytywnej lub negatywnej odpowiedzi zwrotnej na działania agenta);
- wykreowanie agenta przez nadanie jego postępowaniu określonych zasad, reguł i struktur szkoleniowych;
- podjęcie aktywnego „wytrenowania" agenta przez jego zadaniowanie, przyglądanie się ocenie realizacji odpowiedzi na nagrodę i podejmowanie modyfikacji jego działania;
- wprowadzenie agenta do systemu decyzyjnego AML/CFT funkcjonującego w IO.

Należy zauważyć, że ostateczny rezultat takiego postępowania ma pozwolić działać samodzielnie agentowi wobec RL bez ingerencji człowieka. Człowiek występuje w tym przypadku w końcowej fazie decydowania jako np. AMLCO/MLRO (zwłaszcza jako pracownik raportujący do JAF – *reporting officer*) lub też pośrednio jako „czynnik ludzki" podpowiadający agentowi w stanach pośrednich.

W swoim założeniu przedmiotowe opracowanie stanowi propozycję dla decydentów w instytucjach obowiązanych, którzy to decydenci w ramach wykonywania obowiązków ustawowych i wewnętrznych regulacji podejmują działania mające na celu ujawnienie przypadków wykorzystywania tych instytucji w procederze prania pieniędzy czy finansowania terroryzmu. Te działania to procesy decyzyjne i analityczne, zwłaszcza w bankach, które już teraz zachodzą ze wsparciem i z wykorzystaniem sztucznej inteligencji. Propozycja zastosowania metody RL wychodzi naprzeciw dalszym pogłębionym analizom danych, które zwłaszcza instytucja obowiązana uzyskuje w ramach relacji z klientem. W obszarze prezentowanej tematyki jako metody badawcze zastosowano przegląd literatury oraz analizę przepisów prawnych.

Opisywany temat będzie przedstawiony w dwóch częściach, z których pierwsza będzie odnosiła się do struktury samej metody RL i możliwości dostosowania jej do procesu analityczno-decyzyjnego na potrzeby przeciwdziałania praniu pieniędzy oraz finansowaniu terroryzmu z wykorzystaniem pojedynczego agenta. Druga zaś zawiera rozważania co do kwestii związanych z wieloagentowymi rozwiązaniami w środowisku RL oraz z uczeniem się ze wzmocnieniem na podstawie informacji zwrotnej od ludzi, a więc z wykorzystaniem „czynnika ludzkiego" jako podmiotu korygującego.

Całość prezentowanych artykułów ma na celu zwiększenie jakości, efektywności i rzeczowości prowadzonych analiz ryzyka w ramach instytucji obowiązanych, tak aby wygenerowane w ich wyniku zastosowania oraz w konsekwencji użycie środków bezpieczeństwa finansowego pozwoliły na prawidłowe uzyskanie raportów o podejrzanych transakcjach/podejrzanej aktywności, które przekazywane są następnie do Generalnego Inspektora Informacji Finansowej (GIIF) jako do polskiej jednostki analityki finansowej (JAF).

Środowisko, w którym działa agent

Przyjmując, że agent będzie działał w „środowisku" ustalonym przez system przeciwdziałania ML/FT, należy sobie zdać sprawę z tego, że środowiska te będą budowane w specyficzny sposób na potrzeby przestrzegania pewnych ogólnych zasad postępowania wykreowanych w dokumentach, takich jak przykładowo ustawa o p.p.p.f.t., kierunkowe dyrektywy UE/wytyczne FATF (ang. *The Financial Action Task Force*), a także regulaminy wewnętrzne IO. To również polityki (bezpieczeństwa) IO, ale prezentowane w innym wymiarze niż RL, np. uwrażliwienie na obrót gotówkowy, kruszcem, kryptowalutami i posługiwanie się transakcjami finansowymi. Stąd też wielość przeciwdziałań oparta jest na wieloaspektowej ocenie ryzyka instytucjonalnego i indywidualnego, jakie niesie ze sobą klient IO. W tym przypadku zakres środowiska będzie kształtował zarówno klient, jak i transakcje. Ocena zakresu środowiska ma przede wszystkim charakter opisowy odnoszący się także do treści innych dokumentów, jak np. krajowej oceny ryzyka (KOR). Jednym z podstawowych metod postępowania wobec podmiotów IO jest metoda określana jako KYC/KYT (ang. *Know-Your-Customer/Know Your Transaction* – poznaj swojego klienta/poznaj swoją transakcję). Dzięki zastosowaniu ML możliwe jest opracowanie adekwatnych wobec rozpoznawanego środowiska klastrów, przypisując poszczególnym elementom gradację wartości według przyjętego klucza postępowania ocennego (rodzaj czynnika/wielkość czynnika). W związku z tak sformułowanym środowiskiem agent może się w nim poruszać (grać) według ustalonych modeli, polityk lub samodzielnego postępowania (van Keulen, 2021). Taką konstrukcję „środowiska" będzie można odnieść do zbudowania „klastra kont" w IO typu bank. Ten rodzaj klastra będzie odpowiadał postępowaniu klienta (posługiwania się kontem), możliwości obserwowania zasileń konta, redystrybucji środków, konstruowania przepływów w połączeniu z oferowanymi produktami i rodzajem prowadzonej działalności (konto i transakcje będą odzwierciedlały cel aktywności klienta). Tym samym poddany będzie oddziaływaniu agenta nie tylko podmiot – klient (czynnik statyczny), lecz także sama transakcja (czynnik dynamiczny). W przypadku takiego podejścia możliwe jest odkładanie etykiet wobec postępowania klienta jako osoby upoważnionej do dysponowania aktywami, ale także gdy dane do zadania klasyfikacji będą pochodziły

z nieetykietowanych zbiorów uczących, w których brakuje podanych wprost klas do grupowania. Już obecnie w celu wykrycia nietypowych danych za pośrednictwem k-średnich stosuje się tę metodę w celu identyfikacji potencjalnie oszukańczych transakcji wykonywanych kartami kredytowymi czy ujawniania ryzykownego składania wniosków kredytowych (w wyniku ustalania wartości odstających). Wykorzystanie tego typu podejścia do „budowania środowiska" możliwe jest jako grupowanie pojmowane i jako nienadzorowana technika grupowania elementów danych bez wcześniejszej wiedzy o tym, jakie mogą to być grupy. Grupowanie jest zazwyczaj procesem eksploracyjnym. Ale także jako klasyfikacja, czyli technika nadzorowana, która wymaga określenia znanych grup w danych szkoleniowych, w konsekwencji czego każda dana jest umieszczana w jednej z tych grup.

Środowisko AML/CFT jest także pochodną procesów biznesowych czy aktywności w relacji klienta z aktywami, dlatego też należałoby brać w nim pod uwagę również rozwiązania, takie jak modele analityczne (RFM, ang. *Recency – Frequency – Monetary Value* – segmentacja i personalizowanie przekazu marketingowego; LTV, ang. *Loan to Value* – pożyczka od wartości, optymalizator czasu wysyłki, predykcje; NBCh, ang. *Next Best Channel* – model najlepszego następnego kanału), co pozwala badać wartość klienta w czasie, jego potencjał zakupowy, prawdopodobieństwo dokonania transakcji i poziom lojalności (*Lepsze rozumienie klientów dzięki zaawansowanej analityce*, 2023). Takie podejście wynika głównie z tego, że klient instytucji takiej jak bank może być oceniany dwutorowo. Po pierwsze, jako zachowanie się wobec tej instytucji-banku (obsługodawcy), a po drugie, jako utrzymywanie przez klienta relacji z podmiotami trzecimi, co znajduje swoje odzwierciedlenie w „wiedzy" banku jako podmiotu pośredniczącego w tych relacjach. Ta dwutorowość znacznie rozbudowuje środowisko agenta i pozwala na czerpanie wiedzy z różnych podobszarów środowiska, przy utrzymaniu celu dla AML/CFT. Nic nie wyklucza także tego, aby bank skorzystał ze środowiska „niewidocznego" (trzeciego), czyli zewnętrznego, w przypadku którego ocena będzie się dokonywała w relacji czynniki środowiska zewnętrznego – atrybuty wewnętrzne banku wykreowane w relacji z klientem. Sfokusowanie aktywności agenta jednak w tym przypadku zostanie wprowadzone do któregoś z dwóch pierwszych środowisk. Wobec powyższych uwag budowanie ram środowiskowych wydaje się działaniem trudnym i sprowadzającym się do standardu – określonego przepisami oraz oceny indywidualnej – określonej zachowaniem klienta czy podejmowanymi transakcjami.

Zadaniowanie agenta może być realizowane na różnych etapach procesu AML/CFT. Tym sposobem można określić dwa główne podejścia do środowiska AML/CFT, w których będzie mógł działać agent:

Środowisko 1

Jeżeli będzie ono realizowane na etapie początkowym (np. na poziomie identyfikacji i weryfikacji klienta IO) – przy braku (wystarczającej) wiedzy o środowisku, w którym się aktywizuje agent, to jego zadaniem będzie uzyskanie jak najszerszej wiedzy, np. na potrzeby ustalenia skali oceny ryzyka czy zakwalifikowania (kwantyfikacji) klienta, wskazania cech odrzucających akceptację wejścia w relację z klientem przez IO (art. 41 ustawy o p.p.p.f.t.). Prawdopodobnie także ustalenia danych na rzecz wskazania przyszłej polityki dla agenta (ustalenia gradientów nagrody agentowi: za co).

Środowisko 2

Jeżeli zaś agent ma działać na przyszłych zaawansowanych etapach – po podjęciu relacji gospodarczych IO – klient, to jego zadaniem będzie wyśledzenie jak najbardziej adekwatnych, optymalnych schematów postępowania klienta/przeprowadzanych (zleconych) transakcji, które będzie można zakwalifikować jako symptomy podejrzenia prania pieniędzy czy finansowania terroryzmu. Ten zakres środowiska znacznie się poszerzy o inne elementy interakcyjne z agentem, a ponadto prawdopodobnie niezbędne będzie ze względu na wielość scenariuszy wprowadzenie polityki losowej. W tym środowisku polityka powinna być ustalona i podlegać korekcie w zależności od potrzeb.

Ustalanie funkcji nagrody dla agenta w ramach procesu RL/IRL

Z punktu widzenia RL istotnym elementem jest także określenie funkcji nagrody, $R(s, a)$ $R(s, a, s')$. Odgrywa ona zdecydowanie ważną rolę w motywowaniu, a także w sterowaniu agentem. Rzadkie nagrody oznaczają, że agent otrzymuje informację zwrotną dopiero po wykonaniu długiej sekwencji działań, takich jak dotarcie do końca procesu decyzyjnego – wykreowania treści raportu SAR. Opóźnione nagrody zaś oznaczają, że agent otrzymuje informację zwrotną po pewnym czasie, np. ucząc się od nauczyciela-człowieka. Rzadkie i opóźnione nagrody mogą utrudniać agentowi powiązanie swoich działań z wynikami oraz uczenie się metodą „prób i błędów". Dobra funkcja nagradzania powinna odzwierciedlać prawdziwy cel i ograniczenia zadania oraz unikać nagradzania lub karania nieistotnych lub szkodliwych działań (DRF, 2023). Funkcja nagrody, która wykorzystuje modele oparte na fizyce, takie jak cechy geometryczne lub reprezentacje symboliczne, może zmniejszyć złożoność lub niepewność problemu. Ludzka informacja zwrotna oznacza wykorzystywanie bezpośrednich lub pośrednich sygnałów od nauczyciela, ewaluatora lub partnera w celu modyfikacji lub uzupełnienia funkcji nagrody.

W przypadku AML/CFT funkcja nagrody może mieścić się w scedowaniu odpowiedzialności za wykonawstwo z czynnika ludzkiego na nagrodę dla agenta. W konsekwencji system nagradzania może ustanowić stan pożądany przez czynnik ludzki, ale na szerszą i powtarzalną skalę w sytuacji posłużenia się i przetworzenia wielości danych (metadanych) wygenerowanych w IO w ramach realizacji jej codziennej roli wobec produktów i oferowanych usług (odwrotne uczenie się ze wzmocnieniem – ang. *Inverse Reinforcement Learning*, IRL). Chodzi zwłaszcza o uczenie się od czynnika ludzkiego lub od innego agenta. RL – zmierza do maksymalnego zachowania agenta przez ustalone nagrody, IRL – próbuje wydobyć funkcję nagrody z zachowań agenta, dokonując wglądu w jego zachowanie (Alexander, 2018). Tym samym celem jest uzyskanie najlepszej funkcji nagrody przy ujawnieniu optymalnej polityki spośród stosowanych polityk. Istotnym pozostaje zaprojektowanie funkcji nagrody, która stanowi funkcję matematyczną przypisującą wartość liczbową każdemu stanowi lub działaniu agenta (w relacji stan – akcja). Wobec procesu AML/CFT ważny jest również odcinek czasowy aktywności agenta. W przypadku identyfikacji i weryfikacji jako początkowych stanów w relacji do IO odcinek czasowy jest krótki i wymagać będzie takiego dobrania nagród, aby agent w tym krótkim odcinku osiągnął (określony dla niego) cel. Może być to pozytywne skwitowanie „kandydata" na klienta IO i zakwalifikowanie go do dalszych relacji gospodarczych z instytucją, np. przez niestwierdzenie czynników określonych w art. 41 ust. 1 ustawy o p.p.p.f.t., niepotwierdzenie występowania klienta jako podmiotu wyznaczonego na listach sankcyjnych oraz niespełniania warunków deriskingu zgodnie z polityką ustaloną w danej instytucji. W tym przypadku zasób i rodzaj nagród może być przyznawany epizodycznie. Niemniej jednak system AML/CFT jest systemem dynamicznym, co oznacza, że w przypadku ciągłej przestrzeni stanów, jeśli chce się, aby agent łatwo się uczył, funkcja nagrody powinna być ciągła i różniczkowalna. W zakresie stosunków gospodarczych klient – IO, czyli gdy relacje są już nawiązane i nie odrzucone, funkcja nagrody rozciąga się w czasie i może być odpowiednią stymulacją do wyszukiwania negatywnych aktywności opartych na oferowanych produktach i usługach, z których korzysta klient w relacji z IO. Ponadto do projektowania funkcji nagrody należałoby włączyć dokonywanie oceny zachowań klienta jako inicjatora lub beneficjenta „transakcji", z których także należy uzyskać ocenę, czy jest ona (lub może być) ogniwem w łańcuchu procederu ML/FT. Idealnym rozwiązaniem byłaby możliwość ciągu akcji agenta, które działałyby w konfiguracji predykcyjnej, zwłaszcza gdy brak jest wiedzy co do dalszego postępowania klienta lub kolejne ogniwo o_{k+1} nie będzie znane. Na możliwość rozwiązania takiego problemu wskazuje funkcja nagrody w IRL wykreowana w relacji człowiek – komputer. Do wykorzystania pozostają tu nagrody określane jako „rzadka", „gęsta", „ukształtowana", „wewnętrzna" i „zewnętrzna" lub ich konfiguracje. Zbudowanie funkcji nagrody (w ramach RL) nie jest stanem statycznym, to raczej dynamiczny proces ciągłych zmian, doskonalenie, testowanie i ulepszanie wynikające między innymi z iteracji agenta.

Jest to zrozumiałe także ze względu na potrzebę stałego monitorowania klienta/transakcji celem oceny ryzyka indywidualnego i dostosowywania adekwatnego wobec ryzyka pakietu środków bezpieczeństwa finansowego (działanie ciągłe), np. może to wynikać ze zmiany branży klienta, uzyskania kredytu, ustalenia nowych kierunków eksportu towarów, zmiany statusu na relacje gospodarcze z PEP (ang. *Politically Exposed Persons* – osoby na eksponowanych stanowiskach politycznych). Funkcja nagradzania powinna odzwierciedlać prawdziwy cel i ograniczenia zadania oraz unikać nagradzania lub karania nieistotnych lub szkodliwych działań. Generalnym, nadrzędnym celem w AML/CFT jest skuteczne wykrywanie podejrzanych transakcji oraz zlikwidowanie luk wykorzystywanych przez przestępców do prania nielegalnych dochodów lub finansowania działalności terrorystycznej za pośrednictwem systemu finansowego (Kasianova, 2020). Jednakże tego typu cel ogólny składa się z wielości celów pośrednich odnoszących się do poszczególnych klientów czy transakcji. Służą one do niedoprowadzenia do stanu wykorzystania IO przez sprawcę ML/FT lub – w przypadku wystąpienia tego stanu – do zmniejszenia powstałych strat ujawnienia okoliczności, o których informację należałoby przekazać do jednostki analityki finansowej (ograniczenie ryzyka). Zadaniem agenta będzie poszukiwanie stanów okoliczności, które mogą wskazywać na podejrzenie popełnienia przestępstwa prania pieniędzy lub finansowania terroryzmu lub stanów uzasadnionego podejrzenia, że określona transakcja lub określone wartości majątkowe mogą mieć związek z praniem pieniędzy lub finansowaniem terroryzmu. Mając na uwadze różnorodność IO, każda z nich, jeżeli zdecyduje się na użycie agenta w ramach RL, będzie musiała indywidualnie określić nagrodę „za co" i skalę (gradację) nagrody „za coś". Będzie to więc elementem procesu sterowania nie tylko agentem, lecz także realizowaniem taktyki dochodzenia do celu głównego (polityki). Proces AML/CFT jest na tyle dynamiczny, iż będzie on wymagał zmiany nagród w czasie aktywności agenta z możliwością zachowania nagród przy ciągłym wzmocnieniu (nagroda jest uzyskiwana przez agenta za każdym razem, gdy pojawia się pożądane zachowanie). W swojej pracy K. Kasianowa (2020) przyjęła na potrzeby użycia HMM (ang. *Hidden Markov Model*) do AML jako wyróżniki ocenne nagradzania: typ transakcji (podejrzana lub nie) oraz ukrytą (nieobserwowalną) zmienną losową, która jest zależna tylko od poprzedniej wartości (własność Markowa), a także różne cechy transakcji (np. czas transakcji, waluta, kierunki, kwota, liczba transakcji w ciągu ostatnich 7 dni itd.), które posłużyły do definiowania obserwowalnej zmiennej.

Stosowanie nagród, jako podejście epizodyczne, jest możliwe w początkowej fazie relacji klienta z IO, tj. w przypadku identyfikacji i weryfikacji wobec potrzeby stosowania środków bezpieczeństwa finansowego. Należy zauważyć, że w RL to raczej stosowanie nagród niż określona polityka jest tym, co pozostaje istotą dojścia do wyznaczonego celu. W przypadku zadania epizodycznego mamy do czynienia z punktem początkowym i punktem końcowym (stan końcowy). Spowoduje to utworzenie odcinka: listy stanów, akcji, nagród i nowych stanów. Tym samym punktem początkowym jest wola

klienta w kwestii nawiązania relacji z IO, a stanem końcowym decyzja co do możliwości kontynuowania relacji na przyszłość z IO. W konsekwencji, gdy nie nastąpi uzasadniony ze strony IO debanking/derisking i po stronie IO wystąpi wola kontynuowania relacji (np. przez podpisanie umowy o prowadzenie rachunku bankowego), kończy się na tym stosowanie RL wobec potrzeby wstępnego badania ryzyka klienta w IO (np. w ramach KYC, ang. *Know Your Customer* – procedury należytej staranności). W tym przypadku listy stanów pośrednich, potrzeba określonych akcji i przystosowanie nagród są determinowane oceną ryzyka, jakie niesie przyszły klient, zweryfikowaniem jego tożsamości, stanu majątkowego (deklarowanego), biometrii, sprawdzenia w wewnętrznych i zewnętrznych bazach danych śladów jego notyfikacji itp. Proces ten jest do opisania w ramach wewnętrznej procedury (zob. art. 50 ustawy o p.p.p.f.t.) i szczegółowych regulaminów otwierania kont. Przy zadaniu ciągłym – dla agenta zadanie trwa w nieskończoność (brak stanu końcowego). IO realnie nie będzie zainteresowane zakończeniem relacji z klientem. Wyjątkiem będą kwestie wskazane w art. 41 ust 1 i 2 ustawy o p.p.p.f.t. (podobnie gdy wystąpią stany określone w art. 74 i art. 86 ustawy o p.p.p.f.t.). W tym przypadku agent musi nauczyć się wybierać najlepsze działania (optymalne) i jednocześnie wchodzić w interakcje z otoczeniem. W rzeczywistości stan końcowy jest określony, ale nieustalony czasowo. Może być on wynikiem zarówno niewłaściwego zachowania, jak i woli rozwiązania relacji z IO czy zdarzenia w postaci np. śmierci klienta. Można jednak przyjąć, że będzie ono teoretycznie nieskończone na potrzeby budowania akcji i nagród dla agenta. Zadaniem agenta będzie wyszukiwanie takich zdarzeń, które uniemożliwią dalszą relację lub wskażą na podejrzenie udziału w procederze ML/FT (które też można przyjąć jako negatywny warunek rozwiązania relacji). Mając na uwadze to, że w przypadku identyfikacji i weryfikacji odcinek działania będzie dość krótki i skończony, możliwe jest zastosowanie w tym przypadku podejścia metodą Monte Carlo, aby ocenić, jak agent sobie poradził z zadaniem. W przypadku natomiast monitoringu możliwe jest podejście oparte na uczeniu się różnic czasowych (tj. ang. *TD Learning* – uczenie się metodą różnic czasowych). Metoda uczenia ma na celu ustalenie, jak przewidywać wielkość zależną od przyszłych wartości danego sygnału. Ocena funkcji wartości odbywać się będzie po każdym kroku przed kolejnym (t, t+1, R_{t+1} i wartość $V_{(St+1)}$) (Simonini, 2018).

Monte Carlo $$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$$

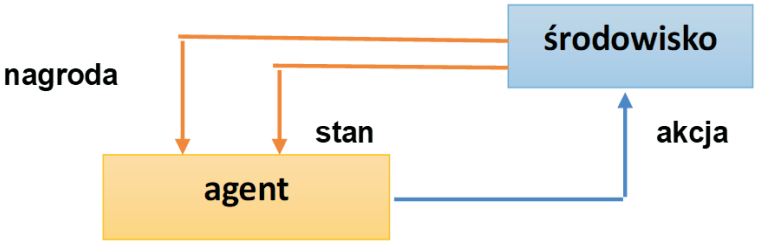
TD Learning $$V(S_t) \leftarrow V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

Problem potrzeby zbudowania „ograniczonego" środowiska dla działania agenta wiąże się głównie z sytuacją w przypadku wstępnych relacji klient – IO oraz gdy mamy do czynienia z krótkimi łańcuchami transakcji przestępczych w procederze ML/FT. Ten drugi problem związany jest również ze specyfiką niektórych IO, jeżeli

rodzaj świadczonych przez nie usług jest „krótki", ale wymagany w samej przestępczej procedurze, np. potrzeby sporządzenia aktu notarialnego w transakcji zbycia nieruchomości. W takim przypadku nadzorca będzie musiał scharakteryzować te czynniki, które w niewielkim przedziale czasowym pozwolą zidentyfikować zagrożenie wobec czynności i czynników ocennych/predykcyjnych (algorytmy wiedzy), jakie powinien wziąć pod uwagę agent, także inicjując dalsze kroki analityczne.

Uczenie przez wzmocnienie z wykorzystaniem pojedynczego agenta – Single-Agent Reinforcement Learning (SARL)

W uczeniu przez wzmocnienie agent wchodzi w interakcję ze środowiskiem przez sterowanie za pomocą trzech sygnałów: stanu x, sterowania (akcji) u oraz nagrody (kosztu sterowania) r. W każdym kroku algorytmu regulator obserwuje stan x_k obiektu, a następnie wykonuje akcję u_k, przeprowadzającą obiekt do następnego stanu x_{k+1}. Jednocześnie regulator otrzymuje sygnał wartościujący wykonaną akcję w postaci nagrody r. Po otrzymaniu nagrody regulator wykonuje kolejny krok algorytmu (Rak, 2013).



Rys. 1. Sposób funkcjonowania RL na rzecz optymalizacji decyzji w środowisku AML/CFT
Źródło: opracowanie własne

Jak można zauważyć (zob. rys. 1), agent (pojedynczy) w każdym kroku postępowania otrzymuje informację zwrotną z otoczenia, następnie wybiera działanie oparte na ustalonej polityce, wchodzi w interakcję z otoczeniem i otrzymuje nagrodę/karę. Obserwacja – działanie (akcja) – nagroda, a w tym stanie powiązań kolejne obserwacje zbierane są w celu utworzenia partii (polityki). Następnie wykorzystywane jest to „doświadczenie" do aktualizacji polityki agenta oraz funkcji wartości z użyciem spadku gradientu. Ten proces jest kontynuowany iteracyjnie, aż do osiągnięcia pożądanej zbieżności lub określonej liczby iteracji (uzyskanie stanu końcowego, zamkniętego). Kwestią więc najważniejszą dla problemu AML/CFT jest to, jakie przyjąć schematy planowania działań, tak aby zakładając dynamiczną strukturę środowiska i różne warianty postępowania, osiągnąć optymalny cel założeniowy.

Funkcje do wykonania jedynie częściowo mogą korzystać z „podpowiedzi" dotychczasowych zewidencjonowanych sposobów postępowania sprawców ML/FT, ale i ze sposobów dochodzenia w IO do wniosków decyzyjnych na potrzeby konstruowania raportów SAR lub innego sposobu informowania jednostek analityki finansowej o swojej podejrzliwości wobec klienta/transakcji. Pozostała część funkcji musi być wygenerowana przez relacje ze środowiskiem i uczenie się, gdyż agent, chcąc osiągnąć cel, musi pozostawać w stałej relacji ze środowiskiem. Jego zasób wiedzy na wejściu musi być uzupełniany jako wynik aktywności zmian środowiskowych będących następstwem np. zmiany produktów, ich modyfikacji (pozostających po stronie IO), a także zmienności taktyk sprawców ML/FT, posługiwania się produktem (pozostających po stronie klienta).

Środowisko, w którym działa agent, jest więc w miarę jasno sformułowane samymi rozwiązaniami prawnymi w zakresie AML/CFT, ale dla większej precyzji powinno być ono indywidualnie dopasowane do określonej IO oraz zakresu świadczonych produktów/usług, a także rodzaju klienta i zlecanych transakcji. Potrzebne więc jest dookreślenie tych elementów środowiska (otoczenia), z którymi w ich masie będzie wchodził w reakcje agent. Nie powinno być to także „środowisko wyjałowione", czyli takie, w którym agent z założenia nigdy nie uzyska sprzężenia niosącego wartość nagrody. Dlatego tak ważne jest dokonanie w IO właściwego grupowania elementów w środowisku na potrzeby RL. W zakresie systemu AML/CFT nadal najważniejsze pozostają na poziomie IO czynności określane jako *onboarding* (identyfikacja i weryfikacja) klienta oraz w przypadku pozytywnej weryfikacji na dalszym etapie powtarzanie tego zadania, a także monitorowanie transakcji (zaawansowanych relacji z IO). Tak więc dalszy etap relacji gospodarczych IO – klient pozwalać powinien na budowanie profilu klienta zarówno z punktu widzenia jego cech indywidualnych uzyskanych w ramach identyfikacji i weryfikacji, jak i cechujących go zachowań transakcyjnych (czynnego monitorowania). Zwłaszcza w tym ostatnim zakresie można byłoby doszukiwać się wykorzystania RL na potrzeby typowania podejrzanych cech klienta. ML w opcjach dla AML/CFT można doszukiwać się tam, gdzie występuje możliwość swobodnego wyboru atrybutów danych (cech podejrzeń), a także wystarczająca dostępność danych wysokiej jakości (na przykład w scenariuszach, w których następuje szybki przepływ środków i można wziąć pod uwagę dużą liczbę zachowań i ich atrybutów, ponadto jakość może wynikać z zauważalnych anomalii ilościowych). ML jest również odpowiednie, gdy trudno jest zidentyfikować dynamikę i relacje między czynnikami ryzyka. Aby jednak skutecznie przeciwdziałać w tym zakresie, IO musi od samego początku procesów związanych ze świadczonymi produktami/usługami wprowadzić reżim pozyskiwania, weryfikacji i takiego sposobu wpisywania danych do wewnętrznych systemów informacyjnych, by zapewnić nie tylko biznesową rozliczalność produktu/usługi, lecz także jej identyfikowalność z punktu widzenia wewnętrznych polityk bezpieczeństwa, naruszeń prawa, zgodności z obowiązkami określonymi

w przepisach AML/CFT, a także w „wykrywczym" procesie decyzyjnym związanym z ujawnianiem niezgodności *compliance* oraz z przepisami prawa karnego (kwestia ta dotyczy głównie struktury danych). Bez skrupulatnej schematyzacji, a dodatkowo zdolności do powiększania zbioru atrybutów i sprzężeń zwrotnych w informacji nie uda się prawidłowo wdrożyć programów opartych na ML. Nie wydaje się także właściwym przyjęcie rozwiązania takiego, iż zachowanie klienta sprawcy pozostaje odmienne wobec klientów nierealizujących przestępczego procederu. Jednym z cech „prania" jest to, aby stworzyć pozory idealnego klienta, legalnych transakcji oraz źródeł aktywów. W takim zakresie cechy zachowań są zbieżne. Stąd RL musi wykorzystywać inne rodzaje cechowań wskazujących na podejrzalność (np. niespójność logowań, atrybuty geograficznego obszaru działania, brak rozliczeń za „posiadane" zaplecze logistyczno-transportowe, częste pomyłki w logowaniu, rozbieżność rozliczeń finansowych na podstawie przedłożonych umów itp.). Dotyczy to nie tylko klasyfikacji danych bieżących, lecz także algorytmów wyszukiwania danych w zbiorach archiwalnych. Jest to konsekwencja tego, że RL wykorzystuje algorytmy, które uczą się na podstawie wyników i decydują, jakie działania należy podjąć w następnej fazie, dążąc do uzyskania lepszej jakości decyzyjnej w dalszej kolejności. Dane budowane „na początku" jako wejściowe do środowiska działania agenta muszą być odpowiednio naznaczone w zakresie ML/FT. Przy czym znając, iż jedną z metod ML jest upodabnianie się do legalnych transakcji, a w zakresie FT jest korzystanie także z legalnych środków na działalność terrorystyczną, naznaczenie musi dotyczyć nie tylko danych wyróżnionych jako świadczących o przestępczym postępowaniu (np. sygnał z listy podmiotów wyznaczonych do list sankcyjnych), lecz także musi dotyczyć to niepodejrzewanych podmiotów, którym produkt/usługę świadczy IO. Tu w grę mogą wchodzić dane, takie jak biometryczne, adresy e-mail, adresy IP, wprowadzenia i wyprowadzenia usługi, skorzystania z usługi nieznanego klienta/jego przedstawiciela (rozporządzenie wykonawcze ministra finansów do ustawy o p.p.p.f.t.), struktura logiczna przesyłanych danych, przestrzeń geolokacyjna (także z uwzględnieniem geograficznego czynnika ryzyka klienta), posiadany status – podejrzewany przez organy ścigania odbiorca płatności itp. Jak można zauważyć w przypadku ML/FT, niekoniecznie dane dotyczące wyłącznie transakcji będą skutecznie wyróżniały przestępczy proceder. Dane mogą być pozyskiwane jako wewnętrzne, tj. z bazy online IO oraz z archiwum czy z baz, z których IO posiada uprawnienia pozyskiwania danych oraz z danych zewnętrznych, jeżeli IO uzyska możliwość ich identyfikacji na potrzeby przeciwdziałania ML/FT.

Dla skuteczności realizacji RL niezbędne jest wykonanie znacznego wysiłku w zakresie identyfikacji i weryfikacji, aby otrzymać jak najbardziej pewne dane wejściowe. Chodzi tu zwłaszcza o uzyskanie wiedzy nadającej się i możliwej do przetworzenia matematycznego na potrzeby posłużenia się tymi danymi na dalszych etapach analizy z wykorzystaniem RL (zbudowanie adekwatnych czujników bezpieczeństwa wobec uznanych zagrożeń dla środowiska, w którym będzie

aktywizował się agent). Przy czym dane te należałoby uznać za „pomocnicze" do weryfikacji kierunku akcji, a nie jako podpowiedź klucza dla agenta. Wartością dodaną jest tu aktywne samouczenie się agenta oparte na naganie i nagrodzie (bazowanie na sztucznej inteligencji agenta). Stąd także musi wystąpić czynnik ocenny dla agenta, który będzie odpowiedzialny za weryfikację efektu jego pracy. Zadaniem czynnika ludzkiego jest wykonanie czynności, które doprowadzą do zwiększenia (wzmocnienia) zdolności agenta do uczenia się (zwiększenia optymalności wyników pozytywnych). Chodzi zwłaszcza o dostarczenie jak najpełniejszej wiedzy o środowisku, w którym ma działać agent. Oznacza to stworzenie bufora jako magazynu danych przechowującego informacje zebrane przez agenta w trakcie uczenia. Dane te następnie wykorzystane są do jego wytrenowania (Miśtak, 2023). RL ma za zadanie także dokonywać powtarzalności pewnych procesów, jednakże znacznie szybciej, niż to wykonywałby inny czynnik, np. ludzki. W konsekwencji chodzi raczej o stan aktywności (dynamiczny), w którym agent uczy się szerokiego zakresu danych ludzkich, a ostatecznie wykonywałby zadania wykraczające poza możliwości ludzkich ekspertów.

Kolejną kwestią jest zapewnienie relacji pomiędzy agentem i środowiskiem. Ten rodzaj relacji określany jest jako pętla uczenia się lub pętla sprzężenia zwrotnego. Agent zostaje wyposażony w możliwość dokonywania określonych akcji (*a*) w środowisku (musi mieć wpływ na środowisko), zwrotnie otrzymuje informację na temat stanu przeprowadzenia akcji. Stan zwrotny stanowi uzupełnienie wiedzy agenta. Jest on także związany z otrzymywaniem informacji o nagrodzie (lub karze). Obydwie te dane zwrotne stanowią wzmocnienie dla agenta, który uczy się i ponawia akcje (*a*') w środowisku. Należy jednak zaznaczyć, iż cele działania agenta w środowisku AML/CFT są przynajmniej dwa główne. Pierwsze to poszerzanie wiedzy – (wykrywanie) nowych schematów postępowania sprawców ML/FT oraz drugie – skupianie się na wykrywaniu rzeczywistych procederów prania pieniędzy czy finansowania terroryzmu. Stąd jawi się przede wszystkim możliwość wykorzystania nadzorowanych algorytmów uczenia maszynowego. Wydaje się jednak, że przy takim podejściu i trzymaniu się jedynie modeli scenariuszy dla agenta nadal nie uzyskuje się właściwego celu, czyli dążenia do uzyskania wiedzy online w czasie rzeczywistym dla przerwania lub dyskretnego monitorowania poczynań sprawcy (możliwe jest jednak określanie prawdopodobnego zachowania się sprawcy).

Ponadto istotna staje się w wielości danych potrzeba ich interpretacji AML/CFT jako niezależnych od osobowego decydenta z wykorzystaniem ML. Czyli bez „podpowiadania" i wygenerowania wyniku jako efektu samodziałań agenta oraz jako efektu samouczenia po procesie wzmocnienia. Należy zaznaczyć, że w systemach AML/CFT występuje cecha niestabilności, która nie pozwala na ścisłe charakteryzowanie elementów ocennych, a także przyjęcie skal nieprawidłowości. Stąd też w RL poważne trudności będzie można spotkać w profilowaniu danych wejściowych i ich wartości, tak aby nie wprowadzać w błąd agenta oraz nie dawać mu do uczenia

danych niesprawdzonych (nawet niemożliwych do zweryfikowania) lub błędnych. Ponadto w takiej sytuacji niekoniecznie możliwe będzie dokonanie prawidłowej oceny „nagrody" na etapie akcji agenta w środowisku.

Dlatego też na potrzeby RL wstępnie istotne byłoby takie podejście, które:

- pozwalałoby, mimo błędnych lub nieweryfikowalnych danych wejściowych, dokonywać z czasem ich walidacji przez samodziałanie (samouczenie) agenta uwzględniające również odrzucenie części danych wejściowych (zwłaszcza początkowych) na dalszych etapach akcji w środowisku i budowanie wiedzy opartej na doświadczeniu samowalidacyjnym i samowykonawczym (eksploracja) lub
- przyjęłoby za „wzorzec" klasyczne wypracowane w IO postępowanie klienta, a zadaniem agenta byłoby jedynie zwracanie uwagi na występujące odchylenia od tego przyjętego modelu na rzecz skumulowania wiedzy o kliencie, która jednak ostatecznie byłaby oceniana przez czynnik ludzki. Zadaniem agenta byłoby więc jedynie uczenie się ze wzmocnieniem dla efektywności wykazywania kolejnych odchyleń od klasycznego wzorca (eksploatacja). Akcja = polityka (stan).

Do zastanowienia pozostaje kwestia, czy agent nie powinien oceniać zwłaszcza danych wejściowych do środowiska ze względu na zbieżność z pierwszym etapem dla procederu prania pieniędzy, jakim jest *placement*, czyli umieszczenie nielegalnie pozyskanych środków oraz danych wyjściowych, które w procederze finansowania terroryzmu przybliżają dystrybucje aktywów do bliskiego otoczenia zamachowców i czynnych terrorystów. Oczywiste jest, że dane wejściowe i wyjściowe dla środowiska stanowią znacznie większy zakres informacyjny, który wytwarzany jest zwłaszcza w ramach „normalnego" wykorzystywania IO przez klientów, a nie w celu przestępczym. Ponadto dane te mogą być częścią środowiska na etapach pośrednich.

Ustanawianie polityk dla agenta

RL jest częścią sztucznej inteligencji. W tej metodzie ujawnia się ona przez zdolność agenta do wybrania strategii zmierzającej optymalnie do wyznaczonego celu. Polityka jest więc strategią, którą agent stosuje w dążeniu do osiągnięcia tych celów. Jednocześnie polityka pozwala zdefiniować mapowanie polegające na tym, że ze zbioru możliwych stanów w środowisku agent ma podjąć właściwe akcje, jeżeli chodzi o nagrody. W konsekwencji możliwe jest określenie polityk, jakimi kierować się będzie agent w środowiskach AML/CFT:

- W środowisku, w którym ustalona jest polityka, dla agenta zadaniem będzie potrzeba prowadzenia polityki optymalnej [stan – działanie – nagroda (+)/ kara – nagroda (-)]. Jest to polityka dana, pasywna. Agent uczy się tylko użyteczności stanów $U\pi(s)$ lub użyteczności par stan – akcja $Q\pi(s, a)$;

– W przypadku gdy polityka nie jest ustalona przez decydenta IO, agent jest podporządkowany polityce dostosowywanej do optymalizacji wyników, samoucząc się kolejnych kroków na dotychczasowych wynikach lub działając samoczynnie z jednoczesnym korzystaniem z podpowiedzi (sterowanie). Aktywne uczenie.

Mając na uwadze to, że nie można w początkowym okresie relacji klient – IO podejmować czynności uznanych za stosowanie środków bezpieczeństwa finansowego z nowym klientem, w celu uzyskania szerokiego spektrum scenariuszy adekwatnych dla danego środowiska klienta, którym posłużyłby się agent, możliwym wyborem pozostaje podjęcie działania opartego na modelu (jeżeli nie można być innowacyjnym, to przynajmniej trzeba być maksymalnie klasycznie wykonawczym – uczenie pasywne). Możliwe jest więc sklasyfikowanie wyjściowo możliwie wszystkich stanów środowiska w celu podjęcia akcji przez agenta lub ustalenie wszystkich możliwych stanów negatywnych i sprawdzenie, czy w tym okresie one występują (przeszukanie środowiska). Podstawowym wyznacznikiem staje się mapowanie polityki o ustalonych (historycznie) stanach działania (stanach znanych dla danego środowiska). Agent nie podejmuje żadnych decyzji, musi robić to, co dyktuje mu (historyczna) polityka. Konstrukcja samego modelu statycznego stanowi kumulację wywnioskowania ocen postępowania wynikającego „z innych polityk" (Foffano, Russo, Proutiere, 2023), np. z doświadczenia, a także z oceny ryzyka instytucjonalnego czy dokumentów, takich jak krajowa ocena ryzyka (KOR), lub doświadczeń innych IO. Dopiero w kolejnych fazach wdrażania scenariuszy można podjąć czynności na potrzeby budowania bardziej rozbudowanych funkcji dla agenta, a zwłaszcza przejścia na RL w celu jego usamodzielnienia zmierzającego do optymalizacji wyników i ustabilizowania polityki wykonawczej akcji dostosowanej do indywidualnych wymagań (budowania kolejnych stanów na podstawie uzyskanych wyników ze stanów poprzednich). Taki funktor dotyczy zakresu identyfikacji i weryfikacji klienta, któremu umożliwia wejście jedynie z wynikiem pozytywnym na poziom relacji gospodarczych z IO (art. 39 ust 1 ustawy o p.p.p.f.t. z wyjątkiem ust. 2). Będzie to stan terminalny kończący dotychczasową politykę i jednocześnie rozpoczynający prowadzenie polityki opartej na aktywnym uczeniu się agenta. Jest to podejście ostrożnościowe ze względu na to, że decydent niekoniecznie będzie miał możliwość na tak wczesnym etapie dokonania ostatecznej weryfikacji klienta, a jedynie uczynienie jej na takim poziomie, iż będzie on pozwalał na kontynuowanie relacji z IO. Decydent na potrzeby późniejszych relacji klient – IO musi znaleźć optymalną politykę π, w której także powinien wykorzystać dotychczasowe osiągnięcia agenta. Stan terminalny będzie stanem początkowym dla dalszych akcji agenta w uczeniu aktywnym. Na drugim etapie relacji IO – klient będą ustalane także stany terminalne dla agenta, jednakże ich cel wobec stanu początkowego będzie odmienny.

Gdy polityka agenta $\pi(s)$ jest z góry ustalona, może on obserwować, co się dzieje, czyli wie, do jakich stanów dociera i jakie otrzymuje w nich nagrody. Stanem

początkowym „s" będzie zdarzenie wyrażenia woli nawiązania trwałych relacji przez klienta z IO, a w wyniku tego uruchomienie czynności formalizujących te relacje. Należy jednak pamiętać, że nagrody otrzymywane w stanach nieterminalnych nie są dla agenta istotnym kryterium – liczy się tylko suma nagród otrzymanych na drodze do stanu terminalnego, zwana wzmocnieniem. Zadaniem agenta jest nauczenie się wartości użyteczności stanów $U^{\pi}(s)$, obliczanych zgodnie z równaniem:

$$U^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R(s_t)\right].$$

Celem działania agenta jest obliczenie użyteczności stanów $U^{\pi}(s)$ związanych z własną polityką $\pi(s)$. Użyteczności stanów zdefiniowane są jako wartości oczekiwane sumy nagród (dyskontowanych) otrzymanych przez agenta startującego z danego stanu i poruszającego się zgodnie ze swoją polityką (zob. wzór równania). W tym przypadku pojęcie użyteczności określa ilość korzyści, którą agent wynosi lub osiąga dla danego wyniku gry (uzyskanych stanów). Przyjęcie polityki działania musi powodować maksymalizowanie się nagród. Agent wykonuje przebiegi uczące (ang. *trials*), w których przeprowadza akcje zgodne z własną polityką, aż do osiągnięcia stanu terminalnego. W każdym kroku otrzymuje percept wskazujący zarówno bieżący stan, jak i nagrodę. Agent może szacować tę wartość, obliczając tzw. nagrodę pozostałą (ang. *reward-to-go*) dla każdego stanu odwiedzonego w danym przebiegu. Na koniec przebiegu agent oblicza nagrodę pozostałą w stanie końcowym jako nagrodę otrzymaną w tym stanie. Następnie, cofając się w przebiegu, oblicza nagrody pozostałe dla wcześniejszych stanów jako sumy nagród otrzymanych na końcowym odcinku przebiegu (Paluszyński, 2012; Jaśkowiak, 2016).

Opis algorytmu opiera się na teorii procesów decyzyjnych Markowa (ang. *Markov Decision Process*, MDP). Proces decyzyjny Markowa (MDP) to matematyczne sformułowanie procesu decyzyjnego. Agent jest decydentem. W ramach uczenia się przez wzmacnianie jest uczniem lub decydentem. Istotne jest w taki sposób przekazać informacje temu agentowi, aby mógł on nauczyć się podejmować decyzje. Jako taki MDP jest krotką: $\langle S, A, P, \gamma, R\rangle$ (stan, akcja, przejścia prawdopodobieństwa, czynnik dyskontowy, nagroda). Tego rodzaju podejście możliwe jest dla ograniczonej i zamkniętej macierzy wyników dla agenta. Stąd możliwe jest jej np. wykorzystanie w fazie początkowej nawiązywania relacji przez klienta i IO, gdy sama IO dysponuje schematami postępowania na potrzeby identyfikacji i weryfikacji „przyszłego klienta". Ale także wówczas, gdy działa oparty na jakiejś ustalonej polityce zastosowania ograniczonej ilości środków bezpieczeństwa finansowego, decydując się jedynie w skrajnych przypadkach na nienawiązywanie relacji z klientem. Te przypadki jako element „prostej" polityki muszą być wykryte przez agenta, np. w zakresie nawiązywania relacji na odległość czy korzystania z bankowości

internetowej. Wydaje się, że takie podejście może być także podejściem weryfikowalnym klienta już w ramach prowadzonych relacji, ale związane z ograniczonym macierzowo i wynikowo zbiorem postępowania, np. gdy klient uzyskał dostęp do nowej, dotychczas nieużytkowanej usługi IO.

Do rozważenia pozostaje także kwestia możliwości badania przez agenta IO klienta jako „innego agenta" w środowisku podejmowanej działalności gospodarczej, tj. czy podejmowane przez niego przedsięwzięcia związane są z rzeczywistym rodzajem działalności nastawionym na zysk (zwiększenie aktywów), czy wytracaniem majątku (osobiste działanie na szkodę firmy) lub jego przekazywanie podmiotom trzecim niezgodnie z jego taktyką intencjonalną gospodarowania nim (wyprowadzanie majątku z firmy w celu przestępczym). Obejmuje to działanie, w którym znajduje się taki układ, aby nagroda nie trafiła do uprawnionej osoby, lecz do jakiejś powiązanej (np. przestępczo, słupa) osoby trzeciej. Takie podejście jest do przyjęcia, gdy sama IO realizuje na zlecenie klienta usługi księgowe czy firma posiada w IO konto, np. w związku z rozliczeniami podatkowymi i prowadzoną działalnością gospodarczą (konto firmowe, mikrorachunek). Każde „odnalezienie" takiego elementu powinno łączyć się z wyznaczeniem nagrody dla agenta. Kumulacją nagrody powinno być dojście do takiego stanu, który będzie można sklasyfikować jako podejrzenie ML/FT (całkowita nagroda). Będzie on wymuszony odnalezieniem pośrednio kilku elementów (pośrednio nagradzanych), z których ten stan będzie się składał. Nie można wykluczyć, że proces ten będzie tylko mógł być częściowo obserwowalny lub też zawierał ukryte (dla agenta) schematy postępowania. Wypracowanie skutecznej polityki agenta będzie polegało na przejściu ze stanu klasycznego/historycznego do stanu aktualnego/inicjacyjnego/predykcyjnego, co prawdopodobnie będzie wymagało przeszacowania macierzy przejść czy wprowadzenia nowych zmiennych lub wprowadzenia dodatkowego agenta (przy jednym obserwatorze). Należy zauważyć, że agent będzie się poruszał w niezwykle dynamicznym środowisku (np. wykonywania zarówno znacznej liczby operacji bankowych, jak i przyjmowania nowych klientów lub rozwiązywania umów z dotychczasowymi). W konsekwencji stale będzie się zmieniała także wartość poszczególnych elementów nagradzanych lub będzie w niektórych przypadkach zanikała (np. przez rezygnację klienta z relacji z bankiem).

Z pewnością będzie mógł on przyjąć jedną z opcji taktycznych: eksploracyjną lub eksploatacyjną. Pierwsza posłuży agentowi do ustalenia nowych dróg dojścia do maksymalnej nagrody i stanu celowego, ale kosztem czasu realizacji. Druga zaś pozwoli na zastosowanie już znanych mu sposobów identyfikacji maksymalizacji stanu, ale nie umożliwi ustalenia możliwych alternatywnych (może lepszych) sposobów dochodzenia do celu. Rozstrzygnięcie w zakresie przyjętej taktyki związane będzie z indywidualną potrzebą przeprowadzenia określonych operacji identyfikujących stany ML/FT. Oznacza to, że albo decydent będzie zadaniował agenta na potrzeby jedynie przeszukania danych (czy występuje powtarzalność pewnych

zachowań klientów w modelu podejrzalności) albo też w kierunku bardziej ambitnym związanym z potrzebą wyszukania nieznanych scenariuszy procederu przestępczego związanego z ML/FT. Przy czym możliwe są następujące scenariusze: poszukiwania schematu przestępczego (modelowego, jeżeli taki zostanie stworzony) lub poszukiwania odchyleń działania klienta (odejście od modelu, ale niebędące przyjętym schematem przestępczym) oraz anomalii zachowań, które jedynie prawdopodobnie będą ustalone jako wynik pozytywny, jednakże bez kwalifikacji ich jako schematu przestępczego. W tym ostatnim przypadku działanie agenta będzie związane z uczeniem się na potrzeby ustalania nowych postępowań klienta wzbudzających podejrzenie. Do rozstrzygnięcia u decydenta pozostanie to, czy za takie „wykrycia" agent byłby wynagradzany równie gradacyjnie co za rzeczywiste wykrycie nieprawidłowości. Po ich wykryciu i zbadaniu będzie je można włączyć do zbioru uczącego, jako „wiedzy" o niepożądanej aktywności klienta, jeżeli tak sklasyfikowane będzie to ustalenie. Wydaje się, że agenta należałoby nauczyć polityki wykrywania podejrzalności w środowisku AML/CFT, mimo że wszystkie okoliczności nie będą znane (ukryte). Kolejne akcje agent podejmuje według strategii π, jednakże strategia powinna uwzględnić co najmniej stany, takie jak te wskazane w art. 35, 41 i 43, 44 i następnych ustawy o p.p.p.f.t. z uwzględnieniem art. 38 w czasie wskazanym w art. 39 ustawy o p.p.p.f.t. i w każdym innym wskazanym za niezbędny.

Jak wskazano, uczenie ze wzmocnieniem RL różni się od uczenia nadzorowanego zwłaszcza tym, że w uczeniu nadzorowanym dane szkoleniowe zawierają klucz odpowiedzi, a więc model jest trenowany z samą poprawną odpowiedzią, podczas gdy w uczeniu ze wzmocnieniem nie ma odpowiedzi i to agent wzmacniający decyduje, co zrobić, aby wykonać optymalnie zlecone zadanie. W przypadku braku zestawu danych szkoleniowych musi uczyć się na swoich doświadczeniach. Chodzi więc o wyuczenie agentów rozwiązywania problemów o stale rosnącym poziomie złożoności w różnych środowiskach (zwłaszcza chodzi o tzw. środowisko dynamicznie inspirowane aktywnością klienta). Uwzględniając, że RL jest modelem uczenia maszynowego, który można określić jako uczenie metodą „prób i błędów", algorytm nie otrzymuje klucza, według którego mógłby przyporządkować określone dane, lecz zestaw pewnych reguł i oczekiwanych rezultatów. Algorytm taki podejmuje więc działania w celu znalezienia „poprawnej odpowiedzi", czyli osiągnięcia pożądanego celu. W modelu uczenia ze wzmocnieniem można posługiwać się systemem nagród, które są przyznawane za prawidłowe działanie algorytmu (Szostek, Bar, Prabucki, Nowakowski, 2022, s. 13). Algorytm uczenia przez wzmacnianie jest w dużym uogólnieniu rekurencyjną procedurą zdobywania wiedzy metodą „prób i błędów".

Anomalie w sensie negatywnym i kierunkowym (ML/FT), jakie mogłyby być związane z pracą agenta, można skoncentrować na samym zachowaniu, cechach, statusie klienta oraz można je także uplasować w algorytm uczenia maszynowego do wyszukiwania najważniejszych transakcji przestępczych. Dodatkowym wsparciem dla agenta byłoby przygotowanie różnych strategii, jakie znane są IO

w zachowaniach klientów podejrzewanych o udział (kreowanie) zachowań ML/FT. Uczenie się ze wzmocnieniem można traktować jako przeszukiwanie przestrzeni możliwych strategii. Jest ono realizowane w sposób powiązany z rzeczywistymi interakcjami ze środowiskiem. Każda kolejna strategia rozważana w trakcie procesu przeszukiwania powstaje przez modyfikację dotychczasowej strategii na podstawie zaobserwowanego doświadczenia stanu, akcji, nagrody, następnego stanu (*Sztuczna inteligencja/SI Moduł 13*, 2023). Stąd też ważne jest to, aby agent zapoznawał się w ramach procesu uczenia się nie tylko z pojedynczymi zdarzeniami, np. umieszczeniem klienta na liście PEP (obwieszczenie MF), umieszczeniem klienta na liście sankcyjnej jako podmiotu wyznaczonego, lecz także ze scenariuszami związanymi ze specyficznie powiązanymi zdarzeniami kreowanymi przez sprawcę-klienta na potrzeby taktyki przestępczej. Stąd też działanie agenta pozostaje wtórnym wobec innych czynności, jakie pierwotnie należałoby zrealizować, jako wypełnienie celu uzyskania optymalnie efektywnego celu zadaniowania agenta. Do takich czynności zaliczymy: uzyskanie maksymalnie potwierdzonych danych identyfikujących i kwantyfikujących klienta (beneficjenta, pełnomocnika itp.), także z wykorzystaniem agregacji danych, wprowadzenie danych do użytkowanego produktu, które pozostają „czynnikami" emitującymi anomalie od przyjętego wzorca postępowania z nim, uszczelnienie i wyostrzenie wskaźników ryzyka, zwłaszcza indywidualnego klienta, zbudowanie bazy alertów opartej na dotychczasowych doświadczeniach i wiedzy, zapewnienie schematów przekazów sprzężeń zwrotnych do punktu kumulacji wiedzy dla agenta oraz zapewnienie dostępu do weryfikowalnych zewnętrznych (wobec IO) baz danych poszerzających wiedzę o platformie funkcjonowania klienta, np. dotyczy to: mediów społecznościowych, rządowych baz danych, danych typu *open source* czy publicznych archiwów.

Uzyskanie właściwego przekazu pomiędzy komórką AML w IO a agentem będzie wymagało zarówno opracowania adekwatnego kanału przekazu danych, a także wzajemnego rozumienia się, zwłaszcza zgodności interpretowania czy zgodności modelowania zachowań klienckich. Rozwiązanie takie może być przynajmniej dwuaspektowe:

- odwzorowanie w zachowaniu klienta schematów uzyskiwanych na bieżąco z różnych źródeł jako potwierdzone zachowanie podejrzane (zindywidualizowane podobieństwo ogólne) lub
- inicjowanie wyjaśnień klienta za pomocą stosowania przez IO adekwatnych do ustalonego ryzyka wywołanego przez klienta/transakcję środków bezpieczeństwa finansowego (podejrzalność indywidualna).

Wydaje się, że przeszukiwanie środowiska dla agenta można sprowadzić jedynie do pewnego zakresu rodzaju opcji. Wymaga to jednak wysiłku ze strony IO, którego celem byłoby przyjęcie na potrzeby uczenia się określonego stanu wiedzy – uzyskania zbioru modeli podejrzanych transakcji. Zadaniem agenta byłoby nie tylko ich stałe poszukiwanie (odwzorowań w czasie rzeczywistym), lecz też zdobywanie

wiedzy o prawdopodobieństwie ich wystąpienia „na przyszłość" na podstawie już posiadanej wiedzy i doświadczenia (wzmocnienia) [*jeżeli w ustalonym wzorze taktyki przestępczej klient-sprawca podejmował określony rodzaj schematu wykorzystania usługi IO (pierwsze działanie), to w dalszej kolejności podejmował innego rodzaju przedsięwzięcia, inicjując innego rodzaju transakcje lub np. wnioskowanie o kredyt (dalsze działania). W konsekwencji, gdy agent stwierdzi stan pierwszego działania, może określić czas i prawdopodobieństwo wystąpienia dalszych działań układających się w taktykę przestępczą. Jego kolejną akcją będzie odnalezienie zachowań podobnych do dalszych działań*]. Początkowo wiedza zgromadzona w komórce AML ma być pomocna do działania agenta i dlatego to ostatecznie przedstawiciel tej komórki musi właściwie zinterpretować dane dostarczone przez agenta w ramach RL. Dotyczy to zarówno aktualnych, jak i przyszłych scenariuszy przeciwdziałania z użyciem RL. W celu uzyskania odpowiedniego efektu niezbędne będzie zapewnienie wstępne zgodności RL z KYC/KYT w ramach AML/CFT.

W uczeniu nadzorowanym możliwe jest więc przyjęcie określonego modelu „wzorcowego postępowania" dla klienta instytucji obowiązanej (IO), a w konsekwencji – w ramach analizy AML/CFT – dokonywania porównania między zachowaniem klienta a przyjętym modelem, czy nie zachodzą odchylenia. Przy czym dla całości działań IO najważniejsze będzie to, czy zachowanie reprezentowane przez klienta kwalifikuje się do podejrzanych okoliczności lub uzasadnionego podejrzenia procederu ML/FT, a nie do nie-przestępczego, hybrydowego postąpienia z produktem/usługą. Stosowanie analizy rozpoznania ML/FT w IO i wsparcie jej ML kończy się na jakimś etapie podjęciem decyzji (np. w przypadku stanu terminalnego kwalifikowanego jako SAR). Decyzja dana nie jest jedyną i może mieć ona kilka wersji. Ich rodzaj zawarty został w ustawie o p.p.p.f.t. Uczenie się przez wzmacnianie RL wykorzystuje algorytmy, które uczą się na podstawie wyników i które decydują, jakie działania należy podjąć w następnej kolejności. Po każdym działaniu algorytm otrzymuje informację zwrotną, która pomaga mu określić, czy dokonany przez niego wybór był prawidłowy, neutralny czy błędny. Jest to dobra technika do stosowania w zautomatyzowanych systemach, które muszą podejmować wiele drobnych decyzji bez przewodnictwa człowieka (Bajaj, 2023). Warto przywołać fakt, że niezależnie od tego, jak IO dochodzi do etapu decyzyjnego związanego z transformacją informacji analitycznej na informację jednostki analityki finansowej (JAF), zawsze ostatecznym decydentem nie jest maszyna, lecz człowiek. RL może więc przybliżyć uzyskanie wysokiego stanu pewności (przez maksymalne zmniejszenie niepewności) związanego ze stanem prawdopodobieństwa oceniającym możliwość wystąpienia okoliczności i zdarzeń świadczących o ML/FT.

Wydaje się, że w ostatnich kilkudziesięciu latach mimo zmienności w czasie środowisko AML/CFT pozostaje jednak w miarę konstruktywnie budowane. Podejmowanych jest ponadto wiele prób wzmocnienia dotychczasowych rozwiązań polegających między innymi na wydawaniu wspólnych wytycznych, które

poszczególne kraje powinny w podobny sposób implementować, oraz prowadzeniu wspólnej polityki bezpieczeństwa UE spełniającej się w zakresie wydawania dyrektyw w obszarze AML/CFT. Dotyczy to także stałego prowadzenia raportów ewaluacyjnych w poszczególnych krajach zwracających uwagę zwłaszcza na potrzebę dochowania zgodności z przepisami UE, rozbudowanie zakresu podmiotowego działań *compliance* zwłaszcza w instytucjach obowiązanych, przygotowanie europakietu przepisów mających na celu zwiększenie efektywności działań w ramach UE w obszarze AML/CFT. Taki stan rzeczy pozwala na w miarę obszerne zebranie instrumentów przeciwdziałania, których użycie ma przede wszystkim rozpoznać, a następnie określić zakres aktywności sprawcy kwalifikowany prawnie i faktycznie jako pranie pieniędzy czy finansowanie terroryzmu. Nie należy także zapominać, że środowisko, w którym będzie działał agent na potrzeby AML/CFT, jest kojarzone z instytucjami obowiązanymi, które jednocześnie kwalifikuje się do instytucji finansowych. One dostarczają i kreują najwięcej informacji na rzecz JAF, a jednocześnie wskazane procedery ML/FT również charakteryzuje posługiwanie się zwłaszcza aktywem finansowym. Stąd też na potrzeby uzyskiwania danych uczących możliwe byłoby korzystanie także z agentów wykonujących zadania w obszarze finansów czy fintech, np. wdrożenia RL w sektorze finansowym polegają na wykorzystaniu botów handlowych, do których jest zaprogramowane uczenie się od środowiska rynku akcji i handlu przez interakcje z rzeczywistym rynkiem, zarządzanie i optymalizacja portfolio, alokacja aktywów, scoring kredytowy, wybieranie najlepszej oferty finansowej.

W konsekwencji aby móc rozpocząć ocenę możliwości zastosowania RL, należałoby przyjąć, iż wszelkie czynności agenta będą podejmowane wobec określonego „obiektu" i w określonym środowisku. Mając jednak na uwadze to, że taki obiekt będzie charakteryzował się wysokim poziomem złożoności, dodatkowym wsparciem będzie określenie go jako „sieci" ustalonych powiązań. Sieci te mogą być opisane jako ważone grafy, w których węzły reprezentują konta i krawędzie stanowiące transakcje między rachunkami, oraz wagi na krawędziach, które reprezentują częstotliwość lub intensywność transakcji. Poszczególne elementy sieci będzie można identyfikować i budować jako relacje (powiązania), np. na podstawie przepływów środków finansowych pomiędzy kontami i wykorzystywaniu poszczególnych instrumentów finansowych, a także relacji IO – klient w ramach KYC, KYT i stosowania środków bezpieczeństwa finansowego (prowadzenia analiz rozpoznania w przypadku niemożliwości zastosowania wymaganych środków bezpieczeństwa finansowego). Zarówno proceder prania pieniędzy, jak i finansowania terroryzmu wiąże się z przepływami środków pieniężnych na dużą skalę przez łańcuchy rachunków bankowych budujące sieci relacji pomiędzy poszczególnymi węzłami.

Determinantami zachowań sprawczych będą więc trzy czynniki podstawowe: klient, transakcja i konto. Każdy z nich jest określony pewnymi cechami, które mogą podlegać grupowaniu. W konsekwencji takie działanie powoduje powstanie zbioru

danych identyfikujących każdy z czynników, który można poddać wnioskowaniu, czy mamy do czynienia z wzorowym postępowaniem, czy z anomalią. I tak anomalią może być zarówno odstępstwo od „normalnego" postępowania, jak i „naśladownictwo" legalnego działania. Na takie postępowanie powinien być uwrażliwiony agent RL. Stąd też można się zastanowić, do jakiego poziomu „optymalności" jest w stanie dojść agent. Wydaje się, że pewniejszym stwierdzeniem jest uzyskanie przez agenta wyniku wystąpienia anomalii niż ustalenia, iż w danym przypadku mamy do czynienia z procederem ML/FT. W tym przypadku nieodzownym będzie czynnik ludzki. Pomocne jest wskazanie, że w przypadku inteligencji technicznej badamy „zapisane zachowanie" (okoliczności), a nie stan woli, jak np. w przypadku wykazania winy (to dzieje się na etapie postępowania karnego). Należy jednak pamiętać, że występujący również algorytm zachłannego podejmowania decyzji całkowicie pomija przyszłe szacowane nagrody, zamiast tego agent zawsze wybiera działanie *a* w bieżącym stanie *s*, które ma najwyższe *Q(s, a)*.

Jak wskazano, RL polega zwłaszcza na uczeniu się „metodą prób i błędów", co w konsekwencji ma powodować powstanie efektu utrwalenia lub zmiany preferencji działań w określonych sytuacjach w zależności od wyniku. Zastanawiające jest to, czy wprawienie w takie działanie agenta możliwe jest przez oferowanie mu do rozwiązania zdarzeń fikcyjnych/pozorowanych, jednakże na tyle zbieżnych z rzeczywistością, aby nie dokonać efektu niewłaściwego nauczenia agenta. To w konsekwencji może powodować błędną reakcję na rzeczywiste problemy. Wydaje się, że pomocniczym elementem będzie wprowadzenie elementu stałości, który będzie gwarantowała budowa normy prawnej przepisu karnego, penalizującego określony rodzaj zachowania. Jak można zauważyć, ustawodawca nie penalizuje całości zachowań przestępczych (okoliczności), a jedynie te, które uzna za najbardziej niebezpieczne dla społeczności, w jakiej funkcjonują. Utrzymanie sposobu zachowania uznanego za niezgodne z normą kwalifikuje ją jako „stały punkt odniesienia" wobec stosowania wzmocnionego uczenia. W takiej sytuacji agentowi pozostaje śledzenie otoczenia na potrzeby subsumpcji zauważonego zachowania wobec normy wskazującej zbieżność ze „wzorem negatywnego" spenalizowanego zachowania. Wobec takiego postępowania możliwe jest uzyskanie =, trafienia, ale również i n-trafienia, które trzeba będzie zakwalifikować do anomalii, hybrydy czy pomyłki. Agent musi eksplorować otoczenie, wykonując sekwencje działań za pomocą określonych strategii (Szymczyk, 2022). Wydaje się jednak, że dla agenta zawężenie środowiska jedynie do ścisłego identyfikowania czynu zabronionego, tak jak go kreuje przepis prawa karnego, będzie zdecydowanym skondensowaniem akcji, a wręcz stanem nieosiągalnym. Akcje agenta nie muszą dotyczyć „winy", lecz podejrzalności, tak jak ją określa przepis art. 74 i 86 ustawy o p.p.p.f.t., a to znacznie szerszy obszar do identyfikacji zachowań podejrzanych. Kwestia istotna to „precyzja" wyniku uzyskanego przez agenta. Niekoniecznie jego prawdopodobieństwo będzie zawsze wysokie. W tym przypadku możemy użyć przybliżenia funkcji, co pozwala nam uwidocznić

Q(s, a) z użyciem różnych innych cech, zamiast przechowywania jednej wartości dla każdej pary: stan i działanie. W ten sposób algorytm jest w stanie rozpoznać, które ruchy są na tyle podobne, że ich oszacowana wartość również powinna być podobna, i wykorzystać tę heurystykę w podejmowaniu decyzji (Yu, Malan, 2023).

IO jest zobowiązana do prowadzenia analizy ryzyka indywidualnego. W konsekwencji realizacji tego obowiązku najistotniejszymi czynnikami ryzyka wydają się być dla IO jej klienci, kraje (siedziby) prowadzenia działalności, rodzaje produktów lub usług oraz kanały dystrybucji. W związku z tym IO musi określić odpowiednie ryzyko profilu każdego ze swoich klientów i transakcji oraz stosować adekwatny poziom kontroli tego ryzyka. Ten rodzaj oceny IO może także profilować sieciowo i otrzymywać indywidualną ocenę sieci powiązań danego klienta, zwłaszcza gdy zdecyduje się na utrzymywanie z nim relacji gospodarczych. Sieć może być pomocna do stałej analizy zachowań klienta, szczególnie gdy zostanie ona sprofilowana przez relacje z innymi podmiotami oraz przez zlecanie transakcji finansowych. Dodatkowo IO przy budowie powiązań klienta może brać pod uwagę inne czynniki, które w ocenie tej instytucji bardziej będą wzmacniały ocenę ryzyka i będą bardziej adekwatne wobec profilu. Ponadto podejście sieciowe może być pomocne w zakresie budowania środowiska, w którym będzie działał agent na zasadzie cechowania węzłów i relacji pomiędzy nimi występujących (obiekt penetracji agenta – sieciowy obraz środowiska). Wydaje się, że w takim przypadku niezbędne byłoby przeprowadzenie badań, czy prawa rządzące sieciami nie mają wpływu negatywnego (ograniczającego) polityki akcji agenta, zwłaszcza z uwagi na relacje pomiędzy węzłami uzyskane ze względu na ich status w sieci.

Dodatkowym elementem budowania sieci może być także przyjęcie cechowania węzłów jako odnoszenie się do jakiegoś przyjętego w IO dla danego produktu/usługi wzorca i efektu wystąpienia anomalii w zachowaniu klienta, które także będzie można włączyć do jego sieciowego profilu. Zastosowanie oceny sieciowej pozwala także na ujawnienie i piktograficzne przekazanie transakcji, jakie IO wytypuje jako SAR. Mając na uwadze potrzebę przygotowania dla agenta skończonej liczby stanów, należałoby sprowadzić ją do określonej liczby kroków wynagradzanych. Tym etapem końcowym byłoby dokonanie oceny co do potrzeby „zamiany" stanu stwierdzonego w SAR (podobnie dla: STR, ang. *Suspicious Transaction Report* – raport o podejrzanych transakcjach; CTR, ang. *Currency Transaction Report* – raport o transakcjach walutowych) i skierowanie go do JAF przez skorzystanie z art. 86 ustawy o p.p.p.p.f.t. czy z art. 41 ust. 2 w związku z ust. 1 tejże ustawy. IO nie może utrzymywać długiego stanu ryzyka/negatywnego prawdopodobieństwa, ponieważ taki stan może przekształcić się w stan niepoprawności, jako niepewność, od którego trzeba odejść już na początku relacji z klientem, lub zagrożenia (oceny przewyższenia apetytu nad ryzyko). Zastosowanie algorytmu RL – polegającego na maksymalizowaniu długoterminowym sumy nagród „r" – nie będzie więc miało na celu utrzymywanie relacji klient – IO, a ich przerwanie. Przy czym do agenta

będzie należało „zebranie dowodów" na okoliczności ML/FT, które poddane będą następnie ocenie AMLCO w kierunku zerwania relacji lub poinformowania o nich JAF. Tym samym IO powinna dążyć do określenia w miarę ścisłych czynników decydujących o stanie końcowym określonym w art. 41 ust. 1 w związku z art. 34 ust. 1 ustawy o p.p.p.f.t. Temu określeniu mogą służyć ustanowione i wdrożone wzory postępowania z produktem/usługą, schematy oceny ryzyka, sprawdzenia w zbiorach własnych i zewnętrznych. O ile można mnożyć działania na etapie analizy klienta, stosowania środków bezpieczeństwa finansowego, to w przypadku RL należałoby ustanowić pewien stan dopuszczalnej wartości oceny klienta, której przekroczenie jest już niedopuszczalną relacją IO – klient. W tym przypadku pomocna może się okazać ocena apetytu na ryzyko indywidualne i możliwość jego tolerowania do jakiejś wyznaczonej granicy, po której nastąpi stan niedopuszczalnego ryzyka ze względu na bezpieczeństwo IO i innych klientów, a także narażenie się na kary administracyjne za nierealizowanie obowiązków AML/CFT. Algorytm RL uwzględnia więc w poszczególnych posunięciach określone „stany", np. w przypadku własnej informacji/danych IO (pochodzącej z archiwum, pisma organów ścigania, zapytania/żądania GIIF, obserwowania posługiwania się kontem przez klienta) lub z „relacji", w tym przypadku oddziaływania wzajemnego IO – klient (też traktowanej jako określony stan). O ile pierwsze stany mogą być z założenia „ukryte", a więc generowane bez wiedzy i kontroli klienta, to w przypadku drugich mogą być także „ukryte" lub prowokowane/inicjowane przez samą IO. W przypadku tej drugiej relacji IO musi sobie zdawać sprawę z tego, że klient będzie miał świadomość, że jest w określony sposób monitorowany. Nie dotyczy to wyjątków, np. klientów o statusie PEP, których z założenia uprzedza się, że podlegają wzmożonym środkom bezpieczeństwa finansowego w ramach utrzymywania relacji z IO, a także samego środowiska PEP (osoby powiązane gospodarczo i rodzinnie). Stan stanowi więc zmienną, która identyfikuje aktualną pozycję agenta w środowisku.

Zastosowanie RL zarówno w „pierwotnym", jak i w „zaawansowanym" środowisku możliwe jest na następujące potrzeby: zdjęcie klienta – (0/1) – rozpoznawalność – dane biometryczne potwierdzające/zaprzeczające, PESEL klienta – zgodność z formułą zapisu cyfrowego – fałszywość dokumentu/niezgodność, zdjęcie budynku klienta – (0/1) – podany adres – zgodność lub nie podanego miejsca zamieszkania, zastosowanie loginów (ale także innych danych identyfikacyjnych dostępu) – potwierdzenie prawidłowości – ujawnienie zwiększonej ścieżki negatywnych prób wejść/zgłoszenie anomalii (taki stan rzeczy może być identyfikowany np. przez posłużenie się bankowością internetową przez innego użytkownika lub możliwością podjęcia próby fraudu na szkodę klienta), behawioralne zachowania klienta – posłużenie się kartą kredytową – preferencje zakupowe, ale także anomalijne nieuprawnione posłużenie się systemem płatności związanym z IO, ustalone miejsce geolokalizacji klienta – dokonanie czynności bankowości internetowej – anomalia co do miejsca przebywania wobec miejsca dokonania zlecenia usługi/płatności, ubezpieczenie na

życie/onkologiczne – sposób dystrybucji środków/cedowanie uprawnień. Ale w grę będą wchodziły także dane, takie jak geograficzne logowania poza krajem, koszyk produktów i taktyka konsumencka klienta.

W celu umożliwienia funkcjonowania agenta w stworzonym środowisku pomocne będzie określanie postępujących po sobie stanów negatywnych, których apogeum będzie dojście do przerwania lub wyeliminowania relacji IO – klient/transakcja (zwłaszcza z zastosowaniem algorytmu *Q-learning*). Mając na uwadze, że agent może działać w myśl określonej (optymalnej) polityki, tym samym optymalną polityką jest po prostu wybranie działania, które maksymalizuje funkcję wartości dla każdego stanu. Musi więc nastąpić wyszukanie takiej funkcji, która zapewni optymalne uzyskanie rezultatu (w przypadku działania pozamodelowego). Stąd też pomocne powinny być instrukcje postępowania (np. przyjęte na podstawie art. 50 ustawy o p.p.p.f.t.) i schematy *compliance* uwzględniające apetyt na ryzyko i przekroczenie jego dopuszczalnej bariery bezpieczeństwa. Istotne pozostaje również to, aby agent wiedział, jak zostały ocenione jego decyzje, co może powodować staranie się o to, w jaki sposób zmodyfikować strategię, aby w kolejnych krokach otrzymać jak najwyższą ocenę. Oceny za podjęte akcje są skutkiem pewnego sprzężenia, które działa na uczący się system (Paluszyński, 2007).

Na potrzeby budowania strategii AML/CFT będzie można wykorzystać inne działania, takie jak personalizacja produktów, wewnętrzne modele zarządzania ryzykiem, rekomendację indywidualną produktów, weryfikację danych identyfikacyjnych klienta, chatbot/e-maile z prowadzonych rozmów. W takim przypadku wyznaczenie jakościowych reguł postępowania w procedurze decyzyjnej AML/CFT powinno pozostać kompatybilne z tym, jaką rolę będzie wykonywał w tym środowisku agent. Temu służy algorytm *Q-learning*, którego zadaniem jest przypisanie wartości parom (stan, akcja), a nie samym stanom. Funkcja, która odwzorowuje parę (stan, akcja) na wartość, nazywana jest funkcją *Q*. Oznacza to, że zamiast poszukiwać funkcji wartości optymalnej polityki, szukamy funkcji *Q* optymalnej polityki (Kaduri, 2021). Sama metoda jest oparta na wartościach dla RL bez modelu. Zrobimy to na przykład na podstawie doświadczenia, które można uzyskać z wykonywania działań na świecie i zbierania nagród. Podejście *Q-learning* ma na celu określenie optymalnego działania opartego na jego aktualnym stanie. Metoda *Q-learning* zamiast użyteczności, uczy reprezentacji działania (akcja) – wartość w postaci funkcji *Q(s, a)*. Ta funkcja wyraża wartość wykonania akcji *a* w stanie *s*. *Q-learning* jest metodą uczenia się podejmowanego poza polityką. W tym przypadku zastosowanie równania Bellmana powodowane jest potrzebą, żeby przy braku danej polityki brać pod uwagę najlepsze możliwe akcje:

$$Q(s,a) = R(s) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a').$$

Model zaczyna się od wszystkich estymowanych wartości równych 0 ($Q(s, a) = 0$ dla wszystkich s, a). Po wykonaniu działania i otrzymaniu nagrody funkcja robi dwie rzeczy: 1) szacuje wartość $Q(s, a)$ na podstawie obecnej nagrody i oczekiwanych przyszłych nagród oraz 2) aktualizuje $Q(s, a)$ do wzięcia pod uwagę zarówno stare oszacowanie, jak i nowe. Takie działanie pozwala na otrzymanie algorytmu, który jest w stanie udoskonalić swoją dotychczasową wiedzę bez rozpoczynania od zera (Yu, Malan, 2023). Możliwa jest również aktualizacja lokalna funkcji Q będąca wariantem metody różnic czasowych i wyrażona poniższym wzorem aktualizacyjnym (wzór na aktualizację wartości po zdobyciu doświadczenia), obliczanym ilekroć akcja a jest wykonywana w stanie s, prowadząc do stanu wynikowego (następny stan s') (opracowanie własne na podstawie Wiatrowska, Cieńciała, 2020; Paluszyński, 2012; Bahatt, 2018):

błąd różnicy czasowej

$$Q(s,a) \leftarrow Q(s,a) + \alpha(R(s) + \gamma \max_{a'} Q(s',a') - Q(s,a)).$$

Zaktualizowana wartość $Q(s, a)$ jest równa poprzedniej wartości $Q(s, a)$ oprócz pewnej wartości aktualizującej. Wartość ta jest określana jako różnica między nową a starą wartością, pomnożona przez współczynnik uczenia się α. Kiedy $\alpha = 1$, nowe oszacowanie po prostu nadpisuje stare. Gdy $\alpha = 0$, szacowana wartość nigdy nie jest aktualizowana. Podnosząc i obniżając α, możemy określić, jak szybko poprzednia wiedza jest aktualizowana o nowe oszacowania (Wiatrowska, Cieńciała, 2020). Główną częścią jest błąd różnicy czasowej. Mierzy on różnicę między obecną wartością Q, jaką otrzymano dla bieżącej pary (stan, akcja), a najwyższą wartością Q następnego stanu z dodaną do niej bieżącą nagrodą. Algorytm ten w środowisku epizodycznym implementowany jest w następujący sposób: należy utworzyć tabelę wartości dla każdego stanu i powiązanych z nim możliwych akcji. Kolejno ustalić liczbę epizodów. Dopóki nie zostanie osiągnięty stan końcowy, należy wybrać akcję wedle obranej strategii, podjąć akcję a i obserwować nagrodę r oraz kolejny stan s'. Następnie zaktualizować wartość oczekiwanej nagrody według wzoru $Q(s, a) + \alpha(r + \gamma * max_a Q(s', a) - Q(s, a))$. Jeżeli został osiągnięty stan końcowy, należałoby przejść do kolejnego epizodu. Przy czym max_a oznacza, iż sprawdzane są wartości, jakie dadzą nam możliwe akcje w następnym stanie, i pod uwagę bierze się tę najwyższą. Algorytm w ten sposób zakłada, że agent zawsze wybierze najbardziej „opłacalną jakościowo" akcję – rodzaj takich algorytmów nazywa się *off-policy*. Co oznacza, że algorytm do obliczenia wartości Q korzysta z wartości, jaką otrzyma w wyniku postępowania wedle strategii zachłannej, tzw. maksymalnej wartości (Wiatrowska, Cieńciała, 2020).

Środowisko jest modelowane jako proces decyzyjny Markowa, który jest procesem stochastycznym w czasie dyskretnym, w którym podstawowym założeniem jest to, że stan środowiska zależy wyłącznie od jego poprzedniego stanu i działań podejmowanych przez agenta. Przy takim podejściu w pierwszej kolejności agent powinien wyznaczyć kompletny model przejść dla wszystkich akcji (zob. algorytm ADP, adaptacyjne programowanie dynamiczne). Następnie należy wyznaczyć politykę optymalną, spełniającą równanie Bellmana, jak w zwykłym problemie decyzyjnym Markowa. Proces decyzyjny Markowa (MDP) pomaga definiować modele dla decyzji sekwencyjnych. Proces decyzyjny Markowa można sformułować za pomocą 5-elementów $\langle S, A, R, p, \gamma \rangle$, gdzie S to zbiór stanów środowiskowych, A to przestrzeń działań agentów, $R : S \times A \times S \rightarrow$ R to nagroda uzyskana przez agenta za przejście do stanu s', wykonując akcję a w stanie s, R jest zbiorem liczb rzeczywistych, $p : S \times A \rightarrow \Delta (S)$ to prawdopodobieństwo przejścia ze stanu $s \in$ S do stanu $s' \in S$, biorąc pod uwagę akcję a, a $\gamma \in [0, 1]$ jest czynnikiem dyskontującym w czasie. Rozwiązanie MDP polega na poznaniu polityki $\pi : S \rightarrow \Delta (A)$, która maksymalizuje oczekiwaną nagrodę w czasie, gdzie $\Delta (\cdot)$ jest prawdopodobieństwem simplex. Metoda *simplex* – polega na zbliżaniu się do rozwiązania optymalnego w kolejnych krokach. Po skończonej liczbie kroków rozwiązanie zostaje znalezione lub okazuje się, że ograniczenia są sprzeczne albo rozwiązanie jest nieograniczone. Funkcje stanu-działania (funkcja Q) i wartości są takie same, gdzie R $(st, at, st{+}1)$ to środowisko natychmiastowej nagrody zwracane, gdy agent wykonuje akcję at w kroku czasowym t, aby nastąpił tranzyt stanu od st do $st{+}1$ (Zhou, Liu, Tang, 2023, s. 3-4).

Podejście *Q-learning* może to osiągnąć przez opracowanie własnego zbioru zasad lub odejście od zalecanej polityki. Wartość Q jest metryką używaną do pomiaru działania w określonym stanie (zob. metoda – różnicy czasowej czy równanie R. Bellmana). Modele *Q-learning* działają na zasadzie prób i błędów, aby nauczyć się optymalnego zachowania dla zadania. Tablica (tabela) Q zawiera kolumny i wiersze z listami nagród uznane za najlepsze działania każdego stanu w określonym środowisku. Tablica Q pomaga agentowi zrozumieć, jakie działania mogą prowadzić do pozytywnych rezultatów w różnych sytuacjach. Wiersze tabeli przedstawiają różne sytuacje, z którymi agent może się spotkać, a kolumny przedstawiają działania, które może podjąć. Gdy agent wchodzi w interakcje z otoczeniem i otrzymuje informacje zwrotne w postaci nagród lub kar, wartości w tabeli Q są aktualizowane, aby odzwierciedlić to, czego nauczył się model (Kerner, 2023).

Jak można zauważyć, uczenie agentów opiera się zwłaszcza na sprzężeniach zwrotnych, dzięki którym można ocenić, czy działanie agenta było właściwe i zasługuje na nagrodę, czy nie oraz czy mieści się w obszarze błędu. Stąd też formuła RL powinna zawierać jak najwięcej potwierdzeń lub zaprzeczeń ocen efektywności działania agenta. Tego typu sprzężenia w skali samej IO mogą okazać się niewystarczające, zwłaszcza ich „moc weryfikacyjna" może okazać się niewielka. Stąd też należałoby poszukiwać rozszerzenia tej formuły *feedback*. Agent uczy się, dostosowując się do

pozytywnych lub negatywnych informacji uzyskiwanych drogą zwrotną. Do tego typu rozwiązań zaliczyć można weryfikację informacji na platformach wsparcia wymiany informacji prowadzonych pomiędzy IO, komunikatów informacyjnych od JAF lub od organów ścigania. Ważne jest więc także uzyskanie zwrotnych potwierdzeń jakości informacji IO od JAF. Uzyskanie potwierdzeń umożliwia modyfikowanie tabeli *Q* dla agenta, a tym samym doskonalenia systemu wykrywczego opartego na RL. Chodzi zwłaszcza o uzyskanie informacji, w wyniku której dokona się wynagrodzenia *r* agenta za pozytywne rozpoznanie sytuacji. Rozszerzeniem RL w środowisku AML/CFT może być także wykorzystanie tej metody w innym środowisku obsługiwanym także przez niektóre IO. Będzie to środowisko biznesowe, a konkretnie będzie ono dotyczyło czasu transakcji, jej wpływu na rynek, ustalania najlepszych możliwych harmonogramów handlowych, określania krótkoterminowych prognoz finansowych czy w zakresie oceny zamówień w przypadku ich zrównoważonej tożsamości.

Należy jednak zwrócić uwagę na to, że uczenie maszynowe mogą tworzyć samonapędzające się pętle sprzężenia zwrotnego, które szybko stają się tak złożone, że ta analiza nie może już wyjaśniać, jak one działają. Samowzmacniające się pętle mogą również rozprzestrzeniać uprzedzenia i błędy. Na przykład boty napędzane sztuczną inteligencją, które automatycznie agregują zawartość kanałów informacyjnych, mogą rozpowszechniać niezweryfikowane informacje i plotki, co pozostaje istotne w ramach sprzężeń odnoszących się do danych zewnętrznych (Canhoto, 2021). W takim przypadku może mieć zastosowanie aproksymacja funkcji polegająca na zapisie badanej funkcji (np. *U*) w postaci nietablicowej, np. wyrażeniu jej jakąś formułą skończoną. Podobnie jak w konstrukcji funkcji heurystycznych, można zastosować liniową kombinację jakichś cech stanu (zwanych również atrybutami stanu) (Paluszyński, 2012):

$$\hat{U}_\theta(s) = \theta_1 f_1(s) + \theta_2 f_2(s) + \ldots + \theta_n f_n(s).$$

Jak zauważa W. Paluszyński, sukces uczenia się ze wzmocnieniem w takich przypadkach zależy od trafnego wybrania funkcji aproksymującej. Jeśli żadna kombinacja wybranych cech nie może dać dobrej strategii gry, to żadna metoda uczenia jej nie wygeneruje. Z kolei wybranie bardzo rozbudowanej funkcji z dużą liczbą cech i współczynników zwiększa szanse na sukces, ale kosztem wolniejszej zbieżności i zarazem wolniejszego procesu uczenia (Paluszyński, 2012). Wydaje się, że ten sposób podejścia do zagadnienia funkcji aproksymacji jest możliwy do zrealizowania w sytuacji analiz AML/CFT w IO. W tym przypadku możliwe byłoby określenie skończonej liczby akcji dla agenta w danym środowisku, jednocześnie zintensyfikowanie czasowego dokładania elementów dla akcji po ich weryfikacji, które mogą świadczyć o przestępczym procederze (zgodnie z trendami). To zadanie pozostaje w gestii wewnętrznej procedury IO.

Jeżeli chodzi o czynniki zewnętrzne, to sama IO powinna określać skończoną liczbę elementów do sprawdzenia. Możliwe jest także rozwiązanie, że dany stan wymaga dokładanego wskazania elementów do sprawdzenia dla akcji agenta. Wydaje się, że IO powinna opracować zbiór „danych stanów aproksymujących" adekwatnych wobec rodzaju świadczonych usług i produktów oraz „trendów przestępczych" charakterystycznych dla tego rodzaju instytucji (analiza ryzyka instytucjonalnego – art. 27 ustawy o p.p.p.f.t.) oraz charakterystycznych dla cech produktów/usług oferowanych dla klienta (analiza ryzyka indywidualnego – art. 33 ustawy o p.p.p.f.t.). Wynikiem będą stany aproksymujące, które można użyć na potrzeby uczenia się wiedzy przez agenta.

Polityka (strategia) optymalna ma tę właściwość, że jakikolwiek jest stan początkowy i początkowa decyzja, pozostałe decyzje muszą tworzyć politykę (strategię) optymalną ze względu na stan wynikły z pierwszej decyzji – zasada optymalności Bellmana, $Q(s, a) = r + g \max (Q(s', a'))$. W konsekwencji zasadę tę będzie można zastosować zarówno w przypadku stanu początkowego, gdy IO będzie zdawało sobie sprawę z tego, że klient może być osobą podejrzewaną/kojarzoną z procederem ML/FT, jak i gdy stan początkowy takiej wiedzy nie przewiduje i w dalszych relacjach z klientem agent będzie starał się „ujawnić" negatywne związki takiego klienta z ML/FT (metody oparte na zasadach dla RL bez modelu). Stąd też podstawową ideą jest rozpoczęcie od jakiejś polityki, próbkowanie z niej odcinka i aktualizowanie zasad zgodnie z tym, jak dobry był odcinek. Trudną rzeczą jest określenie, jak zaktualizować zasady zgodnie z próbkowanym odcinkiem, na co poszukują odpowiedzi metody gradientu polityki (Kaduri, 2021). Polegają one na wyznaczaniu kolejnego kierunku poszukiwań na podstawie znajomości gradientu funkcji celu w punkcie wyznaczonym w poprzednim kroku. Wyszkolenie agenta można oprzeć na „wskazówkach środowiskowych", które otrzymuje się w wyniku wygenerowania wiedzy z samej IO, platformy, do której ona przynależy (np. w ramach grupy), ale także z ogólnych informacji umieszczonych w krajowej ocenie ryzyka, raportach bezpieczeństwa, komunikatach służb z realizacji spraw itp. (zob. art. 36 i art. 37 ustawy o p.p.p.f.t.). Założeniem jest to, aby agent zasilony był także dwoma kierunkami wiedzy:

- danymi i algorytmami taktyki rzeczywistych sprawców przestępstw ML/FT oraz przestępstw źródłowych współwystępujących z ML/FT oraz
- danymi zmienności taktyki przestępczej w czasie charakterystycznej dla budowania „legalności" i „kurtyny niewiedzy" co do dysponowania aktywami na uprawdopodobnienie legalności środków oraz wsparcie działalności terrorystycznej.

Dotyczy to kwestii, takich jak sposób wykorzystania usługi/produktu, jego rodzaju, pozostawiania śladów, w tym śladów transakcyjnych czy lokalizacji miejsc przestępczej działalności. Ponadto „ślady" mogą być rozpoznawane w przedkładanych przez klienta dokumentach, odmienności deklaracji od realizacji działań, rubrykach wypełnianych przez pracowników IO, konfrontacji z bazami danych czy zapytaniami

podmiotów zewnętrznych, procedurach przetargowych, sprawozdawczości księgowej, finansowej, sposobie zakupu akcji w konfrontacji z informacją o jej wartości itp. Nie oznacza to jednak, że agent będzie uczył się na modelu, niemniej może z tego modelu być wygenerowany „wskaźnik" negatywnych ruchów klienta, który będzie umieszczony w sposobie uczenia agenta. Wobec wielości i różnorodności danych IO-nIO możliwe jest wprowadzenie grupowania. Owe grupowanie stanowi nienadzorowane zadanie uczenia się, które pobiera dane wejściowe i organizuje je w grupy, tak aby podobne obiekty znalazły się w tej samej grupie. Ten rodzaj podejścia, zwłaszcza z zastosowaniem *k*-średnich, można byłoby połączyć z oceną węzłów sieci-środowiska. Celem jest wyszukanie przez agenta takiego czynnika negatywnego w środowisku, które zgłębiać będzie wynagradzanie. Dodatkowym efektem będzie nie tylko wyszukanie takiego czynnika jako „inicjacyjnego" lub elementu „łańcucha" przestępczego, lecz także zbudowanie powiązań pomiędzy poszczególnymi ujawnionymi czynnikami negatywnymi wyszukanego postępowania klienta-sprawcy. Działania agenta muszą jednak być realizowane na danych wiarygodnych i dokładnych oraz „wystarczających" na potrzeby jego uczenia. Stąd konieczność po stronie IO intensyfikacji zwłaszcza procesów identyfikacji i weryfikacji danych w ramach stosowania środków bezpieczeństwa finansowego oraz na późniejszym etapie zapewnienia sobie w ramach monitoringu możliwości weryfikacji danych uzyskiwanych w trakcie utrzymywania relacji gospodarczych IO – klient.

Innym, szczególnym kreatorem środowiska, w jakim działałby agent, byłoby wprowadzenie wobec klienta spersonalizowanej bankowości, którą można byłoby objąć AI (dotyczy to np. klienta o statusie PEP lub klienta n-PEP, wobec którego należałoby zastosować wzmożone środki bezpieczeństwa finansowego – zob. art. 43-46 ustawy o p.p.p.f.t.). W takim przypadku ważne byłoby wypracowanie przez IO (bank) wzorca optymalizacji postępowania na rzecz klienta pozwalającego na nastawienia usług na „optymalny uzysk" wartości aktywów. Proces postępowania klienta wynikający ze specjalnie adresowanych relacji w takim przypadku mógłby być kontrolowany przez RL. Stan taki, zarówno w pierwszej, jak i drugiej sytuacji, wymaga włączenia danych wejściowych, które umożliwią stałą aktualizację wag i maksymalizacji dopasowania poszczególnych węzłów sieci do danych przychodzących. Proces ten pozwala także na ulepszanie pierwotnych danych wejściowych w celu poprawy wydajności i jakości usługi/produktu. Na tak konstruowanych usługach/produktach będzie operować także agent RL. Przy czym dane wyjściowe to może być także zestaw transakcji, które są uważane za zgodne z nietypowym wzorcem (np. stronami transakcji). W ramach przyjęcia aktywnego wpływania na środowisko, jeżeli środowisko jest deterministyczne, to charakteryzuje się ono tym, że następny stan środowiska jest całkowicie zdeterminowany przez aktualny stan i akcję wykonywaną przez agenta, odmiennie w środowisku postrzeganym scholastycznie. I tak, scholastyczny to zmienny proces, którego wynik jest losowy i obarczony pewną niepewnością, odwrotność determinizmu.

Wydaje się, że na potrzeby możliwości działania deterministycznego niezbędne jest, aby za każdą akcją przychodziła zwrotna odpowiedź ze środowiska, stąd też osiągnięcie takich relacji możliwe jest przez odpowiednie etykietowanie zachowań klienta w środowisku.

Podejście indywidualne do ryzyka instytucjonalnego oraz związanego z klientem pozwoli IO na uzyskanie charakterystycznych dla niej wzorców zachowań, które będą kojarzone z działalnością nielegalną. Będą to więc formuły zachowań szczególnych dla rodzaju świadczonych produktów/usług, wielkości i rodzaju IO, czynników ryzyka ustalonych przez ustawę o p.p.p.f.t. oraz te, które wyselekcjonowane zostały w ramach już posiadanego doświadczenia i zaistniałych zdarzeń (ale także jako zachowania unikalne ze względu na możliwość utożsamiania się z legalną działalnością i dokonywania metody miksowania środków legalnych i nielegalnych stosowanych wobec technik ich aktywizowania za pomocą produktów/usług IO). Uzyskane profile, jako efekt oceny, powinny stanowić dla IO wyjątek od całościowych zachowań klientów tej instytucji (jako hybryda lub anomalia). To raczej profil dynamiczny niż statyczny, który aby uzyskać właściwy efekt uczący dla agenta w ramach RL, musi być stale aktualizowany. Pod uwagę, jako źródła danych treningowych, należałoby wziąć zwłaszcza przy FT źródła przestępcze środków wprowadzanych do obrotu (te przede wszystkim ustalone w wyniku działań operacyjnych lub analizy kryminalistycznej). Stąd potrzeba pobierania danych poszerzających środowisko IO także o dane pochodzące z n-IO, zwłaszcza co do typowania źródła dla ML (czyn zabroniony) oraz docelowego beneficjenta (gdy chodzi o finansowanie TR).

Działania takie można byłoby brać pod uwagę także przy ML, ale w tym przypadku „źródłem" środków jest każdy czyn zabroniony. Polski ustawodawca nie zawęził pochodzenia środków do jakiegoś zamkniętego katalogu przestępstw. Ewidentnym źródłem informacji będą zapytania organów ścigania co do pozyskiwania informacji o kontach i związanych z nimi transakcjach, zarówno bezpośrednio wynikających z treści prowadzonych czynności operacyjno-śledczych, jak i ustalonych za pośrednictwem Systemu Informacji Finansowej (ustawa o SInF). Podobnie można byłoby skorzystać ze sprzężenia (IO-USC/KAS) zwrotnego wobec informacji organów celno-skarbowych o podejrzeniu popełnienia przestępstwa podatkowego realizowanego z wykorzystaniem kont bankowych. Identyfikacja danych „odchyleniowych" będzie także możliwa przez zidentyfikowanie szczególnych kont (i ich naznaczenie w danych treningowych) np. związanych ze szczególnym handlem, kont „uśpionych", międzynarodowych przekazów środków do krajów wysokiego ryzyka, kont związanych z osobami powiązanymi z PEP, kont, na których następuje szybkie zasilanie i uwalnianie środków, kont zasilanych wpłatami gotówkowymi czy kont w schematach przestępstw VAT-owskich. W konsekwencji można wykorzystać kombinację analizy wzorców, zestaw danych i dopasowywanie kryteriów w celu identyfikacji kont z podejrzanymi wzorami transakcji finansowych.

Dodatkowym elementem wykonawczym powinien być czynnik czasowy skrócony do minimum (np. codziennej weryfikacji) na potrzeby systematycznego, powtarzalnego, optymalizacyjnego weryfikowania danych w krótkim przedziale czasowym. Czynnik czasowy stosowany byłby niezależnie od epizodów, które wymuszają w IO weryfikację używanych środków bezpieczeństwa finansowego. W tej częstotliwości aktywności agenta powinien skutecznie zastępować czynnik ludzki. Tym samym agent RL będzie wyławiał zdarzenia „nietypowe", które przekazane do analizy przez czynnik ludzki będą mogły być ocenione jako „podejrzane" (np. tak jak jest to określone w art. 74 lub art. 86 ustawy o p.p.p.f.t.). Należy zaznaczyć, że głównym założeniem zastosowania RL jest wprowadzenie do procesu wzmocnionego myślenia o takich mechanizmach, które będą przy wysokich liczbach danych wejściowych ograniczały dane wyjściowe. Odmienne podejście jest podejściem standardowym powodującym powstawanie wielości „czerwonych flag", a tym samym wielościowego przekazywania SAR do jednostek analityki finansowej bez uzyskania wysokiego poziomu użyteczności. RL ma wygenerować tylko „optymalne przypadki" anomalii, które będą także wzmocnione wiedzą i działaniami analitycznymi komórek AML/CFT IO na potrzeby wygenerowania wartościowego SAR (Canhoto, 2021, s. 445-446).

Podejmowanie działań na rzecz osiągnięcia równowagi między eksploracją a eksploatacją

Przyjmując za istotę środowiska, iż klient będzie zachowywał się w nim w sposób nieprzewidywalny, tj. nieplanowany przez IO, wydaje się możliwe wprowadzenie potrzeby równowagi między eksploatacją a eksploracją, czyli pomiędzy przewidywalnym możliwym scenariuszem funkcjonowania klienta (wzorem) a poszukiwaniem postępowania nieodwzorowanego. Koncepcja wykorzystania tego, co agent już wie, w porównaniu z badaniem losowej akcji, nazywana jest kompromisem eksploracja – eksploatacja. W tym zakresie wykorzystuje się algorytm Epsilon-Greedy (epsilon--chciwy *Q-learningu*). Istotną zaletą tego algorytmu jest jego zdolność adaptacji. Wydajność algorytmu można dostroić, dostosowując wartość ε, która kontroluje równowagę między eksploracją a eksploatacją. Wyższa wartość ε spowoduje większą eksplorację, potencjalnie prowadząc do odkrycia lepszych działań, podczas gdy niższa wartość skupi się bardziej na eksploatacji, zapewniając częstsze wybieranie najbardziej znanej akcji. Ta elastyczność pozwala programistom i badaczom dostosować wydajność algorytmu do konkretnych potrzeb ich aplikacji (Łakomska, 2023). Eksploatacja ma miejsce, gdy agent zna wszystkie swoje opcje i wybiera najlepszą opcję na podstawie poprzednich wskaźników sukcesu. Z kolei eksploracja to koncepcja, w której agent nie jest świadomy swoich możliwości i próbuje zbadać inne opcje, aby lepiej przewidywać i zdobywać nagrody. Algorytm chciwy epsilon (Epsilon-Greedy) wybiera między eksploracją a eksploatacją, szacując najwyższe

nagrody. Określa optymalne działanie. Wykorzystuje wcześniejszą wiedzę, aby wybrać eksploatację, szuka nowych opcji i wybiera eksplorację. Tym samym algorytm chciwy epsilon polega na badaniu nowych parametrów i wykorzystywaniu już znanych faktów w celu podjęcia lepszej decyzji. W przypadku przyjęcia funkcjonowania pojedynczego agenta na potrzeby procesu AML/CFT będzie on musiał przyjąć dwie polityki: pierwszą związaną z poszukiwaniem wzorców w środowisku, a drugą związaną z poszukiwaniem nowych „wzorców" w celu doskonalenia akcji i zasilania wstępnego zbioru uczącego. Dlatego też należałoby zakładać rywalizację aktywności agenta pomiędzy eksploracją a eksploatacją. Poświęcenie jedynie jednej z nich zdecydowanie więcej czasu (kosztem drugiej) będzie niosło za sobą negatywne konsekwencje. Idąc w kierunku eksploatacji, agent zauważa wzorce w środowisku, kosztem nowych rozwiązań procederu ML/FT, a patrząc na środowisko przez eksplorację, nadmiernie angażuje się w poszukiwanie nowych nieprawidłowości kosztem czasu wyszukiwania tych znanych. Równoważenie umożliwia agentowi wybrać losową akcję z prawdopodobieństwem epsilon i najlepszą akcję zgodnie z jego bieżącym oszacowaniem funkcji wartości z prawdopodobieństwem 1-epsilon. Rozkład epsilon pozwala uniknąć problemu nadmiernej eksploracji, w przypadku której agent marnuje zbyt wiele czasu i zasobów na badanie działań, które prawdopodobnie nie przyniosą korzyści. Wybranie odpowiedniej strategii dla algorytmu zachłannego epsilon zależy od kilku czynników, takich jak charakterystyka środowiska, cele agenta i dostępne zasoby obliczeniowe (*Epsilon-greedy algorithm*, 2023; Tashmit, 2023). Ponadto należy zwrócić uwagę na to, że epsilon może być stały lub zmienny. Wobec jego zmienności zaproponowano dodatkowo inne rozwiązania, takie jak ε-first czy Value-Difference Based Exploration (VDBE, eksplorację opartą na różnicach wartości), a także adaptacyjną implementację ε-Greedy, w przypadku której nowa wartość parametru ε jest uzyskiwana po każdym kroku uczenia za pomocą równań opartych na rozkładzie Boltzmanna oszacowań funkcji wartości (Mignon, Rocha, 2017, s. 1146-1151). Podejście takie jest ważne, ponieważ występuje w działaniach eksploracyjnych, gdy wiedza o środowisku jest niepewna, na co wskazują zmienne wartości w trakcie uczenia.

Podsumowanie

Wsparcie procesu decyzyjnego w zakresie AMLCFT w IO za pomocą metody RL wymaga przygotowania całości systemu na potrzeby zbudowania środowiska, zbiorów uczących, a także uzyskania czynników wsparcia dla agenta. Niemniej jednak w IO generujących systematyczne metadane wsparcie procesów decyzyjnych za pośrednictwem AI staje się procesem niezbędnym. Tym samym wprowadzenie wsparcia za pośrednictwem metody RL nie jest zadaniem niemożliwym, a wręcz istotnym z punktu widzenia osiągania celów i realizacji obowiązków przez

IO ustalonych w przepisach ustawy o p.p.p.f.t. Prezentowana metoda powinna sprawdzać się w środowisku dynamicznym, co wymaga jednak stałej kontroli agenta i budowania adekwatnego wobec zmienności środowiska systemu nagroda/kara. Początkowy rozwój RL wymaga stałego doskonalenia, zwłaszcza gdy ma być zastosowany w środowiskach niejednorodnych i zmiennych w czasie. Do takich należy środowisko ML/FT. Prace doskonalące RL zmierzają do coraz głębszego zaangażowania w „myślenie techniczne" agentów przez odpowiedni dobór nagród i zdolności adaptacyjnych.

Druga część artykułu będzie poświęcona rozważaniom na temat możliwości wykorzystania więcej niż jednego agenta w środowisku RL oraz realizacji aktywności agenta w ramach uczenia ze wzmocnieniem na podstawie informacji zwrotnej od „czynnika ludzkiego". To podejście umożliwi ocenę wykorzystania agentów w bardziej skomplikowanych i wielowarstwowych środowiskach oraz szybszą korektę funkcjonowania agenta od tej, która byłaby wyłącznie efektem stanu uzyskanego na podstawie własnej „technicznej" oceny uzyskanej w wyniku aktywności.

BIBLIOGRAFIA

[1] Alexander, J., 2018. *Learning from humans: what is inverse reinforcement learning? The Gradient*, https://thegradient.pub/learning-from-humans-what-is-inverse-reinforcement-learning/ (dostęp: 10.09.2023).

[2] Bajaj, P., 2023. *Reinforcement learning*, https://www.geeksforgeeks.org/what-is-reinforcement-learning/ (dostęp: 10.09.2023).

[3] Canhoto, A.I., 2021. Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective, *Journal of Business Research*, nr 131.

[4] DRF, 2023. *What are the best practices for designing reward functions in reinforcement learning for robotics?*, https://www.linkedin.com/advice/0/what-best-practices-designing-reward-functions (dostęp: 5.09.2023).

[5] *Epsilon-greedy algorithm*, 2023. *What are the benefits and drawbacks of using a decaying epsilon strategy in epsilon-greedy algorithm?*, https://www.linkedin.com/advice/0/what-benefits-drawbacks-using-decaying-epsilon (dostęp: 12.09.2023).

[6] Figielska, E., 2011. Ewolucyjne metody uczenia ukrytych modeli Markowa, *Zeszyty Naukowe Warszawskiej Wyższej Szkoły Informatyki*, nr 5.

[7] Foffano, D., Russo, A., Proutiere, A., 2023. *Conformal Off-Policy Evaluation in Markov Decision Processes*, https://arxiv.org/pdf/2304.02574.pdf (dostęp: 12.09.2023).

[8] Jaśkowiak, W., 2016. *Uczenie ze wzmocnieniem na podstawie: AIMA ch21*, http://www.cs.put.poznan.pl/wjaskowski/pub/teaching/wmio/lectures/Uczenie_sie_ze_Wzmocnieniem.pdf (dostęp: 10.09.2023).

[9] Kaduri, O., 2021. *From A* to MARL (Part 5-Multi-Agent Reinforcement Learning)*, https://omrikaduri.github.io/2021/08/07/Part-5-MARL.html (dostęp: 10.09.2023).

[10] Kasianova, K., 2020. *Detecting money laundering using hidden Markov model*, https://dspace.ut.ee/bitstream/handle/10062/68089/kasianova_kseniia.pdf?sequence=1&isAllowed=y (dostęp: 15.09.2023).

[11] Kerner, S.M., 2023. *Definition Q-learning*, https://www.techtarget.com/searchenterpriseai/definition/Q-learning (dostęp: 12.09.2023).

[12] *Lepsze rozumienie klientów dzięki zaawansowanej analityce*, 2023. Persooa, https://www.persooa.com/oferta/zaawansowana-analityka/ (dostęp: 17.09.2023).

[13] Łakomska, D., 2023. *Algorytm epsilona-chciwego: prosta i wydajna metoda eksploracji i eksploatacji*, https://mundurowe.info/wiadomosci/algorytm-epsilona-chciwego-prosta-i-wydajna-metoda-eksploracji-i-eksploatacji/28588/ (dostęp: 12.09.2023).

[14] Mignon, A.S., Rocha, L.R.A., 2017. An Adaptive Implementation of ε-Greedy in Reinforcement Learning, *Procedia Computer Science*, vol. 109.

[15] Miśtak, S., 2023. *Podział modeli uczenia maszynowego wraz z przykładami zastosowania*, https://www.gov.pl/web/popcwsparcie/podzial-modeli-uczenia-maszynowego-wraz-z-przykladami-zastosowania (dostęp: 18.08.2023).

[16] Obwieszczenie MF – obwieszczenie Ministra Finansów z dnia 14 lipca 2023 r. w sprawie ogłoszenia jednolitego tekstu rozporządzenia Ministra Finansów, Funduszy i Polityki Regionalnej w sprawie wykazu krajowych stanowisk i funkcji publicznych będących eksponowanymi stanowiskami politycznymi (Dz.U. 2023 poz. 1632).

[17] Ouyang, L., Wu, J., Jiang, X. et al., 2022. *Training language models to follow instructions with human feedback*, https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf (dostęp: 18.08.2023).

[18] Paluszyński, W., 2007 (wykonał J. Kędzierski). *Algorytm uczenia ze wzmocnieniem dla małego robota mobilnego klasy (2,0). Metody i algorytmy sztucznej inteligencji*, https://kcir.pwr.edu.pl/~witold/aiarr/2007_projekty/roboty2/ (dostęp: 4.09.2023).

[19] Paluszyński, W., 2012. *Uczenie przez wzmocnienie*, https://kcir.pwr.edu.pl/~witold/ai/ml_rl_s.pdf (dostęp: 10.08.2023).

[20] Rak, A., 2013. Zastosowanie algorytmów uczenia przez wzmacnianie w układzie wyznaczania trajektorii zadanej manewrującego statku, *Zeszyty Naukowe Akademii Morskiej w Gdyni*, nr 78.

[21] Rozporządzenie MFFP – rozporządzenie Ministra Finansów, Funduszy i Polityki Regionalnej z dnia 27 lipca 2021 r. w sprawie wykazu krajowych stanowisk i funkcji publicznych będących eksponowanymi stanowiskami politycznymi (Dz.U. 2021 poz. 1381).

[22] Simonini, T., 2018. *An introduction to Reinforcement Learning*, https://www.freecodecamp.org/news/an-introduction-to-reinforcement-learning-4339519de419/ (dostęp: 15.08.2023).

[23] Szostek, D., Bar, G., Prabucki, R.T., Nowakowski, M., 2022. *Zastosowanie sztucznej inteligencji w bankowości – szanse oraz zagrożenia. Analiza prawno-regulacyjna wpływu technologii uczenia maszynowego i pokrewnych na obowiązki sektora bankowego z zakresu zapewnienia zgodności (compliance) oraz zarządzania ryzykiem*, https://us.edu.pl/wp-content/uploads/pliki/PAB_WIB_Zastosowanie_sztucznej_inteligencji_w_bankowosci_Szostek.pdf (dostęp: 2.08.2023).

[24] *Sztuczna inteligencja/SI Moduł 13*, 2023. *Uczenie się ze wzmocnieniem*, https://wazniak.mimuw.edu.pl/index.php?title=Sztuczna_inteligencja/SI_Moduł_13_-_Uczenie_się_ze_wzmocnieniem (dostęp: 2.09.2023).

[25] Szymczyk, K., 2022. *Modern Reinforcement Learning approach and its application*, https://sii.pl/blog/en/modern-reinforcement-learning-approach-and-its-applications/?category=hard-development&tag=embedded-en (dostęp: 12.07.2023).

[26] Tashmit, 2023. *Epsilon Greedy Algorithm*, https://www.codingninjas.com/studio/library/epsilon-greedy-algorithm (dostęp: 5.09.2023).

[27] Ustawa o p.p.p.f.t. – ustawa z dnia 1 marca 2018 r. o przeciwdziałaniu praniu pieniędzy oraz finansowaniu terroryzmu (Dz.U. 2023 poz. 1124).

[28] Ustawa o SInF – ustawa z dnia 1 grudnia 2022 r. o Systemie Informacji Finansowej (Dz.U. 2023 poz. 180).

[29] van Keulen, I., 2021. *Hiding Money Laundering with an Intelligent Multi-Agent System Simulation*, Master's Thesis Artificial Intelligence Department of Artificial Intelligence University of Groningen, The Netherlands, https://fse.studenttheses.ub.rug.nl/25680/1/Master_thesis_Ingeborg_van_Keulen.pdf (dostęp: 9.09.2023).

[30] Wiatrowska, I., Cieńciała, A., 2020. *Q-Learning i SARSA – bez ryzyka nie ma zabawy*, http://main.p.lodz.pl/news.php?id=124 (dostęp: 12.09.2023).

[31] Yu, B., Malan, D.J., 2023. *CS50's Introduction to Artificial Intelligence with Python*, https://cs50.harvard.edu/ai/2020/notes/4/ (dostęp: 10.09.2023).

[32] Zhou, Z., Liu, G., Tang, Y., 2023. Multi-Agent Reinforcement Learning: Methods, Applications, Visionary Prospects, and Challenges, *Computer Science*, vol. 1, nr 1.