**Wyższa Szkoła Zarządzania i Bankowości**

**w Krakowie**

***Symulacja ruchu drogowego z zastosowaniem algorytmów optymalizacji sterowania sygnalizacją świetlną.***

ITS (Intelligent transportation system)  
RL

Actor-Critic (A2C),

SUMO (Simulation of Urban Mobility)

Deep Neural Network, DNN, sieci głebokie

δ, Δ – delta

θ,w – theta, parametry sieci neuronowej (tensora)

* **Wstęp**

Ruch drogowy odgrywa kluczową rolę w funkcjonowaniu wysoko zurbanizowanych społeczności, stanowiąc fundament ich gospodarki i życia społecznego. Dynamiczny wzrost liczby pojazdów wywiera ciągłą presję na istniejącą infrastrukturę transportową. Kongestia drogowa generują wymierne straty finansowe, przyczynia się do zwiększonej emisji spalin, pogarsza jakość środowiska. W sposób pośredni utrudnia i obniżają poziom życia mieszkańców aglomeracji. Zatory drogowe wpływają na czas reakcji służb ratunkowych (straż pożarna, policja, służba zdrowia) co może mieć tragiczne konsekwencje. Wraz ze wzrostem obciążenia infrastruktury drogowej, rośnie zapotrzebowanie na efektywne metody kontroli ruchu. Ponieważ fizyczna rozbudowy dróg, jest bardzo kosztowna, a często niemożliwa, jednym z kluczowych narzędzi poprawy dynamiki ruchu są sygnalizatory świetlne a ich optymalizacja jest kluczowa dla minimalizowania opóźnień drogowych.

Nowoczesne systemy transportowe (ITS), oferują szereg usprawnień podnoszących płynność ruchu drogowego w porównaniu do systemów statycznych, nie uwzględniających dynamicznie zmieniających się warunków środowiska. Systemy takie jak SCATS, SCOOT czy RHODES, pozwalają na adaptacyjne dostosowywanie cykli sygnałów do bieżących warunków drogowych. Mimo ich skuteczności, wciąż istnieje przestrzeń do ich udoskonaleń. W tym kontekście, modele uczenia maszynowego mogą odegrać kluczową rolę w dalszym rozwoju tych systemów.

Nagroda Nobla z dziedzinie fizyki w 2024 roku, jest dowodem, że badania nad algorytmami sztucznej inteligencji pozostają w centrum zainteresowana świata nauki. John J. Hopfield i Geoffrey E. Hinton otrzymali to najwyższe naukowe wyróżnienie za „*fundamentalne odkrycia i wynalazki umożliwiające uczenie maszynowe przy użyciu sztucznych sieci neuronowych*”. Ich prace przyczyniły się do stworzenia mechanizmu wstecznej propagacja błędów, co dało impuls do rozwinięcie neuronowych sieci wielowarstwowych, które stały się podwaliną współczesnych systemów uczenia maszynowego.

Sukces finansowy takich projektów jak CHAT GPT, AlphaFold, Tesla Autopilot, powoduje że ta dyscyplina wiedzy przeżywa kolejny renesans.

* **Uzasadnienie wyboru tematu**

Obecnie jesteśmy świadkami rewolucji AI. Powstają nowe typy jednostek obliczeniowych TPU v6 o prędkości 1836 TOPS (Tera Operations Per Second). Rozwój technologii AI zaczyna być blokowany przez ograniczoną ilość sklasyfikowanych danych niezbędnych do trenowania modeli.   
 Systemy takie jak AlphaGo**,** opracowane przez DeepMind, uświadamiają nam że maszyny mogą przekroczyć poziom ludzkich umiejętności. System **AlphaGo Zero**, osiągnął po 3 godzinach treningu mistrzowski poziom w grze w Go, a po 70 godzinach nauki zaproponował rozwiązania przekraczające dotychczasowe ludzkie doświadczenie.  
 Nowoczesne systemy sterowania ruchem, w połączeniu z technologią autonomicznych pojazdów, mogą znacząco poprawić efektywność komunikacji drogowej. Informacje generowane przez autonomiczne pojazdy oraz inne efektory, mogą stanowić źródło danych do tworzenia zaawansowanych strategii zarządzania ruchem, zwiększając płynność i bezpieczeństwo na drogach.

Wybór tematu pracy jest uzasadniony aktualnymi kierunkami badań w dziedzinie sztucznej inteligencji, potencjałem technologii sieci neuronowych oraz próbą wykorzystania wiedzy z zakresu algorytmów uczenia maszynowego w praktycznym zastosowaniu.

* **Cel pracy**

Celem pracy jest zbadanie, w jaki sposób algorytmy RL, takie jak **aktor-krytyk** **(**Actor-Critic (AC)), mogą zostać wykorzystane do sterowania sygnalizacją świetlną na obszarach o dużym natężeniu ruchu. Symulacje przeprowadzone w środowisku **SUMO** pozwolą na ocenę potencjału oraz efektywności takiego rozwiązania.

Zakres pracy obejmuje:

* **Przedmiotowy**: Optymalizacja sterowania sygnalizacją świetlną na skrzyżowaniach przy użyciu algorytmu aktor-krytyk.   
  Analiza zgromadzonych danych wyników symulacji.
* **Czasowy**:   
  - Analizę literatury i istniejących rozwiązań semestr 5.  
  - Projektowanie i implementację algorytmu semestr 6.  
  - Testowanie i analizę wyników w środowisku symulacyjnym SUMO semestr 7.
* **Przestrzenny**: Symulacje zostanie przeprowadzona w wirtualnym środowisku SUMO. Ruch drogowy będzie generowany w sposób syntetyczny,   
  z uwzględnieniem scenariuszy, które koncentrują się na tworzeniu zatorów drogowych.

**Stosowane metody**

W pracy zastosowane zostaną następujące metody:

* **Analiza źródeł**: Przegląd istniejących systemów sterowania ruchem oraz prac naukowych związanych z zastosowaniem sztucznej inteligencji w tej dziedzinie.
* **Modelowanie i symulacja**: Implementacja algorytmu AC w środowisku SUMO, pozwalająca na symulację sterowania sygnalizacją świetlną w dynamicznych warunkach ruchu.
* **Metody oceny efektywności**: Analiza wyników symulacji, w tym pomiar opóźnień, czasu oczekiwania pojazdów oraz przepustowości, ilość zużytego paliwa i wyemitowanego CO2.

**Opis zawartości poszczególnych rozdziałów pracy**

**Rozdział 4**: Sterowanie ruchem świetlnym, omówienie aktualnych metod sterowania sygnalizacją świetlną.

**Rozdział 5**: Uczenie maszynowe; analiza literatury naukowej, opis procesów RL, AC.

**Rozdział 6**: Pakiet symulatora SUMO.

**Rozdział 7**: Aplikacja algorytmu AC.

**Rozdział 8**: Analiza zgromadzonych danych.

**Podsumowanie**

Praca stanowi połączenie teorii algorytmów sztucznej inteligencji z praktycznym ich zastosowaniem. Celem jest implementacja algorytmu **AC** do sterowania sygnalizacją świetlną, w modelowanym środowisku SUMO.

Przeprowadzone prace i symulacje stanowią doświadczenie edukacyjne, wprowadzające w tematykę algorytmów uczenia ze wzmacnianiem, modelowania systemów transportowych oraz wykorzystania intermodalnego symulatora ruchu drogowego SUMO.

Uzyskane wnioski mogą stać się podstawą dla dalszego pogłębiania wiedzy w zakresie ITS.

* **Sterowanie ruchem świetlnym**

Pierwsze zastosowanie sygnalizacji świetlnej w sterowaniu ruchem drogowym miało miejsce w 1868 roku w Londynie. Latarnie wyposażone były w lampy gazowe. Elektryczna sygnalizacja została po raz pierwszy zastosowana w 1914 roku w Cleveland. Do roku 1918 sygnalizatory były dwukolorowe, tj. wyposażone w światło czerwone i zielone. Trójkolorową, sygnalizacja zawierającą również światło żółte, zainicjował Londynie.

Sterowanie sygnalizacją ewoluowało od systemów stałoczasowych do systemów zmiennoczasowych. Systemy stałoczasowe działałaką na podstawie historycznych danych, bez sprzężenia zwrotnego, zmiennoczasowe dopasowują długość faz lub zmieniając sekwencje faz sygnalizacji do parametrów ruchu.

Nowoczesne systemy obejmują nie tylko pojedyncze skrzyżowania, ale także całe sieci drogowe. Lokalne sterowniki świetlne, działające w zdecentralizowany sposób, są wystarczające w warunkach niskiego ruchu, jednak przy większej gęstości ich wydajność jest niewystarczająca. Skuteczność lokalnych decyzji nie zawsze przekłada się na globalną optymalizację. Obecne trendy to tworzenie scentralizowanych i hierarchicznych systemów sterowania, uwzględniających współpracę między skrzyżowaniami.  
 Najnowsze metody, oparte na modelach predykcyjnych, nie tylko dopasowują sterowanie do bieżących warunków, ale także starają się przewidywać przyszłe sytuacje, co pozwala na lepsze planowanie i podejmowanie decyzji.

**Sterowanie Ruchem Drogowym: Szczegółowy Podział Systemów.**

Poniżej przedstawiono podział systemów sterowania ruchem drogowym.

**1. Podział według struktury sterowania:**

**Systemy zdecentralizowane:**

Lokalne sterowniki sterują ruchem na pojedynczym skrzyżowaniu.

Brak koordynacji między skrzyżowaniami, co ogranicza ich skuteczność w zarządzaniu ruchem na dużych obszarach.

Zastosowanie: Mniejsze miasta lub obszary o niskim natężeniu ruchu, gdzie nie jest konieczna synchronizacja sygnalizacji.

**Systemy scentralizowane:**

Zarządzanie ruchem z jednego centralnego punktu, gdzie zbierane i analizowane są dane z całej sieci drogowej.

Centralny komputer optymalizuje sygnalizację świetlną w czasie rzeczywistym, synchronizując działanie wielu skrzyżowań.

Zalety: Globalna optymalizacja, efektywne zarządzanie ruchem w skali całej sieci.

Wady: Wysokie wymagania infrastrukturalne i obliczeniowe.

**Systemy hierarchiczne:**

Struktura wielopoziomowa, w której każdy poziom odpowiada za inne aspekty sterowania ruchem.

Przykład: Lokalny poziom zarządza sygnalizacją na pojedynczych skrzyżowaniach, a poziom nadrzędny koordynuje większe obszary.

Zastosowanie: Rozległe sieci miejskie z różnymi poziomami złożoności ruchu.

**2. Podział według rodzaju sterowania:**

**Stałoczasowe systemy sterowania:**

* Działają w oparciu o ustalone cykle sygnałów świetlnych, niezależne od aktualnego natężenia ruchu.
* Zalety: Prostota implementacji i niski koszt wdrożenia.
* Wady: Brak elastyczności, szczególnie w warunkach zmiennego ruchu.

**Zmiennoczasowe systemy sterowania:**

* **Systemy akomodacyjne:**
* Zmienna długość faz sygnalizacji bez zmiany ich kolejności.
* Dostosowują się do lokalnych warunków ruchu, ale nie synchronizują z innymi skrzyżowaniami.
* **Systemy adaptacyjne:**
* Dynamicznie dostosowują zarówno długość, jak i sekwencję faz sygnalizacji.
* Wykorzystują dane z czujników w czasie rzeczywistym, co pozwala na optymalizację w zmieniających się warunkach.
* SCATS: System stosowany w Sydney, który dynamicznie dostosowuje sygnalizację w oparciu o lokalne dane ruchowe.
* SCOOT: System używany w Wielkiej Brytanii, optymalizujący sygnalizację w czasie rzeczywistym na podstawie prognoz ruchu.

**3. Podział według technologii i metod działania:**

**Systemy heurystyczne:**

* Wykorzystują reguły oparte na doświadczeniu lub wcześniej zdefiniowane algorytmy zarządzania ruchem.
* Zalety: Łatwe do implementacji i zrozumienia.
* Wady: Ograniczone możliwości optymalizacji w złożonych warunkach ruchu.

**Systemy optymalizacyjne:**

* Stosują modele matematyczne i algorytmy optymalizacyjne, takie jak programowanie dynamiczne, algorytmy genetyczne czy metody Monte Carlo.
* Mogą uwzględniać różne kryteria optymalizacji, np. minimalizację opóźnień, długości kolejek czy emisji spalin.

**Systemy bazujące na sztucznej inteligencji:**

Wykorzystują metody uczenia maszynowego, takie jak:

* **Uczenie przez wzmacnianie (Reinforcement Learning, RL):**  
  Algorytmy uczą się optymalnych strategii sterowania na podstawie interakcji z rzeczywistym środowiskiem.
* **Sieci neuronowe:**  
  Pozwalają na analizę złożonych zależności w danych o ruchu drogowym.
* Programowanie dynamiczne.

Przykłady:

* Systemy adaptacyjne wykorzystujące algorytmy Q-learning lub Advantage Actor-Critic (A2C).
* Metody multi-agentowe, gdzie różne skrzyżowania zarządzane są przez współpracujących agentów.

**Krótki opis działających systemów sterowania ruchem**

* **Urban Traffic Control System (UTCS) to inicjatywa Departamentu Transportu USA,** rozwijana od lat 70. XX wieku, obejmująca cztery generacje strategii sterowania ruchem drogowym:

Pierwsza generacja: Oparta na historycznych danych o ruchu, z planami sterowania zmienianymi co 15 minut.

Czwarta generacja: Oparta na aktualizacjach w czasie rzeczywistym, obliczając moment zmiany fazy sygnalizacji w każdym cyklu.

Ewolucja strategii zmierzała od statycznego do dynamicznego dostosowywania sterowania ruchem, umożliwiając lepszą reakcję na bieżące warunki ruchowe.

* **SCATS (Sydney Coordinated Adaptive Traffic System):**

SCATS (Sydney Coordinated Adaptive Traffic System), opracowany przez australijskich naukowców, to adaptacyjny system sterowania ruchem zaliczany do metod trzeciej generacji. W przeciwieństwie do SCOOT, SCATS nie korzysta z modelu ruchu ani optymalizatora planów sterowania, ale wybiera najlepszy plan sterowania na podstawie bieżących warunków ruchu. Struktura systemu jest hierarchiczna, obejmując trzy poziomy: lokalne sterowniki, urządzenia regionalne oraz centralne centrum sterowania odpowiedzialne za monitorowanie całego systemu.

SCATS dostosowuje długość cyklu, split i offset sygnałów świetlnych, wykorzystując dane z detektorów. Zmiany parametrów, takie jak długość sygnału zielonego, odbywają się w małych krokach co ±6 sekund, co pozwala na dynamiczną adaptację do warunków ruchu. System stosuje mechanizmy głosowania do wyboru odpowiednich planów offsetów i koordynacji między podsystemami.

SCATS jest stosowany w wielu miastach, w tym w Polsce, gdzie został wdrożony w Rzeszowie, Łodzi i Olsztynie.

* **SCOOT (Split Cycle Offset Optimization Technique):**SCOOT (Split Cycle and Offset Optimization Technique) to metoda sterowania ruchem czwartej generacji, zaprojektowana do dynamicznej optymalizacji sygnalizacji świetlnej w oparciu o aktualne dane o ruchu. W systemie tym skrzyżowania są grupowane w pod obszary, a sterowniki w każdym pod obszarze operują na wspólnym cyklu. System dokonuje częstych, niewielkich zmian parametrów, takich jak długość sygnałów, czas trwania faz i offset, w celu minimalizacji opóźnień i zatrzymań.  
  SCOOT korzysta z trzech procedur optymalizacyjnych:

**Optymalizatora splitów**, który analizuje czas sygnałów czerwonych i zielonych, dostosowując ich długość w krokach co 1-4 sekundy.

**Optymalizatora długości cyklu**, który raz na 5 minut zmienia czas cyklu w zależności od nasycenia skrzyżowań w regionie.

**Optymalizatora offsetu**, pracującego raz na cykl dla każdego skrzyżowania, w celu zapewnienia płynności ruchu.

System jest szeroko stosowany w Wielkiej Brytanii i na świecie, a jego najnowsza wersja, SCOOT MC3, wprowadza priorytety dla autobusów i inne udoskonalenia​.

* **RHODES (Real-Time Hierarchical Optimized Distributed Effective System):**

Hierarchiczny system sterowania, który dynamicznie dostosowuje sygnalizację w czasie rzeczywistym, wykorzystując dane z czujników.  
Algorytm ten został nazwany sterowaną optymalizacją faz (COP – Controlled Optimization of Phases). Podobnie jak systemy **DYPIC PRODYN, OPAC** jest oparty na metodzie programowania dynamicznego.

* **GASCAP, SPPORT** Sterowanie ruchem drogowym z wykorzystaniem logiki rozmytej opiera się na analizie długości kolejek i napływu ruchu, które są przekształcane na wartości przynależności do zbiorów rozmytych, takich jak Krótka, Średnia czy Długa. Decyzje sterujące, np. przedłużenie fazy zielonej, wynikają z reguł rozmytych, które uwzględniają siłę aktywacji (FS) dla każdego przypadku. Zaletą logiki rozmytej jest niski koszt obliczeniowy i zdolność lepszego odzwierciedlenia aktualnych warunków ruchu w porównaniu do metod stałoczasowych czy zmiennoczasowych. Przykładowo, długość kolejki o wartości 7 może należeć jednocześnie do zbiorów Średnia i Długa z przynależnością 0,6, co zwiększa możliwości generalizacji. Dzięki temu logika rozmyta jest skuteczną i elastyczną metodą sterowania ruchem drogowym.
* **PIACON** to metoda inteligentnego sterowania ruchem drogowym, opracowana w 2008 roku przez AGH i holenderskiego producenta sterowników, wdrożona w Lubinie. Bazuje na systemach ekspertowych oraz algorytmach optymalizacyjnych i działa na trzech poziomach: lokalnym, arterialnym i sieciowym. Wykorzystując dane z detektorów ruchu, takie jak liczba pojazdów czy długość kolejek. Uwzględnia wielokryterialne podejście, analizując m.in. straty czasu, zatory i emisję zanieczyszczeń, by dynamicznie dostosowywać sygnalizację świetlną do aktualnych warunków drogowych.
* **Systemy oparte na AI:**
* **DRL (Deep Reinforcement Learning):** Wykorzystywane do sterowania sygnalizacją świetlną w oparciu o rzeczywiste dane ruchowe.
* **Metody multi-agentowe:** Agenci zarządzający poszczególnymi skrzyżowaniami uczą się współpracy w celu optymalizacji globalnego ruchu.

Metody adaptacyjnego sterowania ruchem często mają złożoną hierarchiczną budowę i wymagają skomplikowanych algorytmów o dużej złożoności czasowej. Systemy takie jak SCATS i SCOOT są rozwijane i skutecznie sterują ruchem w miejskich sieciach liczących tysiące skrzyżowań. Obecnie dąży się do tworzenia systemów zdolnych do przetwarzania dużych ilości danych w krótkim czasie uwzględniających nietypowe sytuacje takiej jak kolizje czy remonty.

Z badań i wdrożeń przeprowadzonych w różnych aglomeracjach wynika, że zastosowanie zaawansowanych systemów zarządzania ruchem jest korzystne zarówno dla kierowców, pieszych, jak i środowiska naturalnego.

Nowoczesny i wydajny system sterowania ruchem to dziś;

**Skrócenie czasu przejazdu** - optymalizacja pozwalająca zmniejszyć czas potrzebny na pokonanie określonej trasy.

**Większa płynność ruchu** – lepsze zarządzanie eliminuje korki i zmniejsza liczbę nadmiarowych zatrzymań pojazdów.

**Zwiększenie bezpieczeństwa** – poprawa organizacji ruchu zmniejsza ryzyko kolizji i wypadków drogowych.

**Monitorowanie ruchu** - podejmowanie szybkich decyzji w razie utrudnień w ruchu.

**Priorytetowanie pojazdów uprzywilejowanych i komunikacji zbiorowej** – ma szczególne znaczenie dla służb ratunkowych i komunikacji miejskiej.

**Ograniczenie zużycia paliwa i emisji spalin** – zmniejszenie zapotrzebowanie na paliwo oraz ograniczają emisję zanieczyszczeń do atmosfery.

**Personalizowane, planowanie tras** – Dzięki bieżącym informacjom o ruchu, dostarczanym za pośrednictwem znaków zmiennej treści (VMS) czy aplikacji internetowych, użytkownicy mogą wybierać optymalne trasy.

**Dostęp do danych statystycznych** – umożliwiają gromadzenie danych o ruchu w celu dalszych analiz i usprawnień.

* **Uczenie maszynowe**

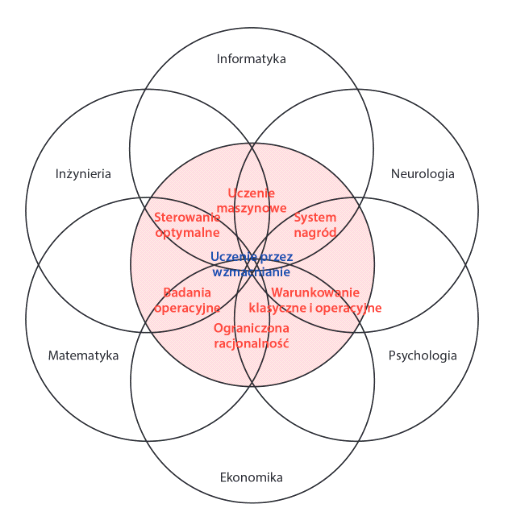
Sztuczna inteligencja (AI), uczenie maszynowe (ML) to dynamicznie rozwijające się dziedziny, które odgrywają kluczową rolę w dzisiejszym świecie technologii informatycznych. Za ojca sztucznej inteligencji i informatyki uznaje się Alana Turing, który w 1943 roku postawił fundamentalne pytanie: "Czy maszyny mogą myśleć?". Jego prace nad maszynami obliczeniowymi zapoczątkowały ideę tworzenia inteligentnych systemów informatycznych.   
Kilka lat później, w 1956 roku, John McCarthy ukuł termin "sztuczna inteligencja" podczas legendarnej konferencji w Dartmouth College, która formalnie rozpoczęła badania nad AI.

W 1959 Arthura Samuela wprowadził termin uczenie maszynowe (machine learning) w kontekście programowania komputerów zdolnych do uczenia się na podstawie danych. Samuel jest również autorem pierwszego samodzielnie uczącego się systemu, programu grającego w warcaby.

Uczenie maszynowe aktualnie dzieli się na trzy główne typy; uczenie nadzorowane, uczenie bez nadzoru oraz uczenie ze wzmocnieniem. W uczeniu nadzorowanym model uczy się na danych z oznaczonymi etykietami, co pozwala na realizację zadań takich jak klasyfikacja czy regresja. W uczeniu bez nadzoru system analizuje nieoznakowane dane, odkrywając ukryte wzorce, na przykład poprzez klasteryzację lub redukcję wymiarowości. Natomiast uczenie ze wzmocnieniem polega na interakcji modelu z otoczeniem, gdzie agent uczy się podejmować decyzje optymalizujące przyszłe nagrody.

**5.1. Wprowadzenie do uczenia ze wzmocnieniem (RL)**

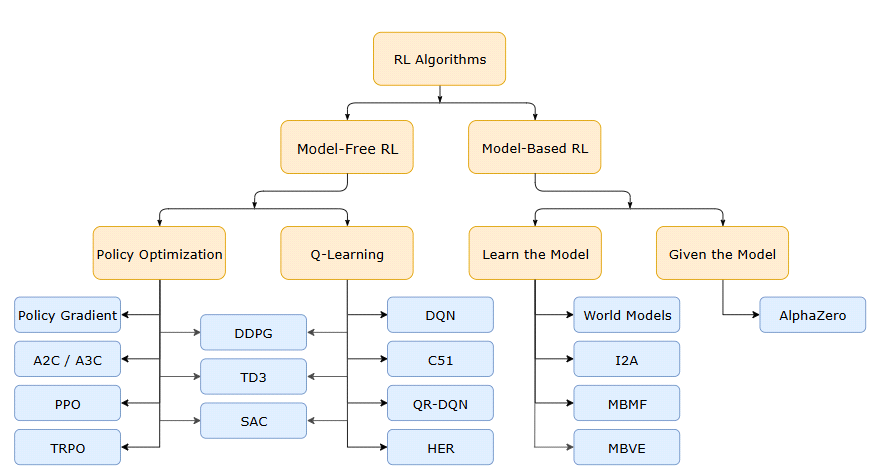
Uczenie ze wzmacnianiem to rodzaj technik stosowanych w systemach uczących się, w których agent podejmuje działania prowadzące do zmaksymalizowania nagrody płynącej ze środowiska, poprzez wykonywanie określonej sekwencji kroków.

Początki uczenia przez wzmacnianie sięgają lat 50. XX wieku. Są silnie zakorzenione w badaniach nad zachowaniem adaptacyjnym, dynamicznym programowaniem i Procesami Decyzyjnymi Markowa. Istnieje wiele obszarów, które są związane z uczeniem przez wzmacnianie. Najistotniejsze przedstawione są na rysunku 1.

*Rysunek 1 - Źródło: schemat pochodzi z książki Głębokie uczenie przez wzmacnianie. Praca z chatbotami oraz robotyka, optymalizacja dyskretna i automatyzacja sieciowa w praktyce. S.31.*

Podstawowy model RL (Reinforcement Learning) wykazuje liczne analogie do modeli psychologicznych z dziedziny warunkowania klasycznego. Eksperymenty przeprowadzone przez Iwana Pawłowa z psami demonstrują zdolność zwierząt do kojarzenia sygnałów środowiskowych, takich jak dźwięk dzwonka, z bodźcami nagradzającymi, np. jedzeniem. Pawłow określił ten mechanizm jako 'wzmocnienie', odnosząc się do bodźca nagradzającego, który wzmacniał pożądane zachowania psa (agenta).

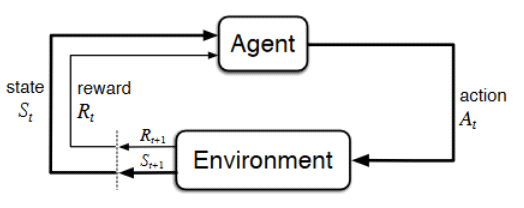
Istnieje dużo algorytmów tego modelu, ale szczególną popularność zyskały obecnie 2 z nich: sieć deep-Q (deep Q-network, DQN) oraz deep deterministic policy gradient (DDPG). Oba są łatwe do wdrożenia, a jednocześnie oferują bardzo duże możliwości adaptacji do śroowiska.

 Na rysunku 2 znajduje się taksometria współczesnych algorytmów RL, zaproponowana przez Josha Achiama, naukowca z OpenAI. Diagram daje pogląd na rozległość tej dziedziny.

**5. 2 Formalne podstawy i terminologia**

Głównymi elementy uczenia przez wzmacnianie są; agent (Agent) i środowisko (Enviroment), kanały interakcji — akcje (action), nagrody (reward) i stany (state).

**Rysunek 1. Schemat blokowy algorytmu RL**

*Rysunek 3,* *Żródło: Schemat pochodzi z ksiązki „Reinforcement Learning: An Introduction” Second edition, in progres November 5, 2017, stron 38*

**Agent i środowisko**

Agent to podmiot, który wchodzi w interakcję ze środowiskiem w dyskretnych krokach czasowych *t,* agent znajduje się w stanie *st∈S ,* gdzie *S* jest zbiorem wszystkich możliwych stanów środowiska*.* W każdym kroku *t* agent wykonuje akcję *at∈A*, odbiera obserwację stanu *st+1* oraz otrzymuje nagrodę *rt+1∈R*, gdzie *A* jest zbiorem dostępnych akcji, a *R* zbiorem możliwych nagród.   
Środowisko reprezentuje wszystko, co otacza agenta, dostarcza mu informacji st+1 i reagując na jego działania at.

**Akcje**

Akcje to działania, jakie agent może wykonywać w środowisku, np. ruchy w grze. Decyzje podejmowane mogą być dyskretne (np. ruch w lewo) lub ciągłe (ustaw czas świecenia światła zielonego na sygnalizatorze na [10,60] s).  
Akcje są częścią **trajektorii**, czyli sekwencji stanów, akcji i nagród, którą agent generuje podczas eksploracji środowiska. Trajektoria zaczyna się od początkowego stanu i kończy się, gdy agent osiągnie stan końcowy lub gdy **epizod** zostanie przerwany po ustalonej liczbie kroków.

**Obserwacje**

Obserwacje to informacje przekazywane agentowi przez środowisko, opisują aktualny stan. Mogą być użyteczne do przewidywania przyszłych nagród.

**Nagroda**

Nagroda w uczeniu przez wzmacnianie to skalarna wartość, którą agent okresowo otrzymuje ze środowiska jako informację zwrotną o jakości swoich działań. Może być pozytywna lub negatywna, ale zawsze ma charakter lokalny, odzwierciedlając niedawne działania agenta, a nie całokształt jego sukcesów. Celem nagrody jest wzmocnienie pożądanych zachowań agenta.

Nagrody pozostają kluczowym elementem procesu uczenia, napędzającym postępy agenta.

**5.3 Procesy Decyzyjne Markowa (MDP)**

Procesy Decyzyjne Markowa (MDP) to model matematycznym używany w uczeniu przez wzmacnianie. Umożliwia formalne modelowanie środowiska oraz interakcji środowiska z agentem. Jest on rozszerzeniem *klasycznego procesu Markowa* dodając do niego terminy akcja i nagroda.

MDP można zdefiniować jako 5-eleentową krotkę:

*MDP = (S,A,P,R,γ) (wzór 1)*

*gdzie:  
S: zbiór stanów środowiska,  
A: zbiór działań agenta,  
P(s′∣s,a) : prawdopodobieństwo przejścia z s do s' po wykonaniu akcji a,  
R(s,a): funkcja nagród, określająca wartość nagrody dla stanu s i akcji a,   
γ∈[0,1): współczynnik dyskontowania, który kontroluje znaczenie przyszłych nagród.*

MDP opisuje, jak działania agenta wpływają na zmiany stanu środowiska oraz na otrzymywane nagrody. Kluczowe na tym etapie są dwie funkcje:

**5.4. Funkcja przejścia *P(s′|s, a)*:**

Funkcja ta definiuje prawdopodobieństwo, przejścia do stanu (s′) po wykonaniu akcji (a) w stanie (s):

(wzór 2)

Funkcja przejścia opisuje dynamikę środowiska oraz określenie wpływu działań agenta na przyszłe stany.

**5.4 Funkcja nagrody *R(s,a)*:**

Funkcja nagrody *R(s,a)* określa oczekiwaną wartość nagrody rt+1, którą agent otrzymuje po podjęciu akcji (*a*) w stanie (*s*). Jest to wartość średnia, uwzględniająca wszystkie możliwe wyniki (nagrody), jakie mogą wystąpić w przyszłości po tej decyzji.

(wzór 3)

*gdzie   
 [⋅]: Operator wartości oczekiwanej, obliczający średnią ważoną wszystkich możliwych wyników.*

Nagroda jest kluczowym elementem kierującym działaniami agenta, ponieważ określa, które stany i akcje są pożądane.

**5.5 Współczynnik dyskontowania nagród γ (gamma).**

Współczynnik określa, jak bardzo agent ceni przyszłe nagrody w porównaniu z bieżącymi. Jeśli γ jest bliskie 0, agent skupia się na natychmiastowych nagrodach, ignorując długoterminowe konsekwencje. Gdy γ jest bliskie 1, przyszłe nagrody są równie ważne jak bieżące, co pozwala na bardziej strategiczne podejmowanie decyzji.”

Agent wybiera akcje tak, aby zmaksymalizować skumulowaną zdyskontowaną nagrodę (G) otrzymywaną w przyszłości. Skumulowana nagroda (lub zdyskontowany zwrot) jest definiowana jako:  
 (wzór 4 można pominąć)

*gdzie:  
Gt: skumulowana zdyskontowana nagroda począwszy od chwili t,  
Rt+k+1R: nagroda otrzymana w kroku t+k+1t+k+1t+k+1,  
γ: współczynnik dyskontowania, który zmniejsza znaczenie nagród otrzymanych w odległej przyszłości.*

**5.6 Polityka.**

Polityka (π) definiuje sposób, w jaki agent podejmuje decyzje w środowisku. Jest to funkcja określająca prawdopodobieństwo wyboru akcji (*a)* w stanie (*s)*:

(wzór 5)

Polityka określa strategię agenta, wpływając na osiąganie celu: maksymalizację skumulowanej nagrody. Polityka optymalna prowadzi do maksymalizacji oczekiwanej skumulowanej nagrody w długim horyzoncie czasowym.

Polityka może być;

- Stochastyczna: Losowy wybór akcji z przypisanymi prawdopodobieństwami, np. eksploracja środowiska.  
- Deterministyczna: Zawsze wybiera tę samą akcję w danym stanie (π(a∣s)=1).

**5.7 Równania Bellmana**

Równania Bellmana są wykorzystywane do rekurencyjnego wyznaczania wartości stanu (V(s)) lub optymalnej polityki (π∗(s)) w danym stanie (s). Ich uniwersalność polega na możliwości zastosowania w różnych technikach optymalizacyjnych, takich jak iteracja wartości, iteracja polityki czy Q-Learning.

**Równanie Bellmana dla wartości stanu (Vπ(s)):**

Wzór na oczekiwana suma zdyskontowanych nagród, zaczynając od stanu s i postępując zgodnie z polityką π.

(wzór 6)

*gdzie*

*Vπ(s) - wartość stanu s przy danej polityce π.  
Ea∼π,s′∼P[.] oczekiwanie (średnia wartość) po losowych zmiennych:  
 a∼π Akcjaa jest wybierana zgodnie z polityką π(a∣s), czyli prawdopodobieństwem   
 wybrania akcji a w stanie s.  
 s′∼P : Nowy stan s′ jest losowany z rozkładu P(s′∣s,a), który opisuje przejścia między   
 stanami w środowisku.  
r(s,a) - Nagroda natychmiastowa za wykonanie akcji a w stanie s.  
γ: Współczynnik dyskontowania (0≤γ≤1).  
Vπ(s′) - Wartość stanu s′, do którego przechodzi system po wykonaniu akcji aaa.*

**Równanie Bellmana dla optymalnej wartości stanu (V∗(s)):**

(wzór 7)

Określa maksymalną możliwą wartość stanu s, gdy agent działa w sposób optymalny.

W przeciwieństwie do wersji on-policy, tu dodany jest operator max⁡\maxmax, który reprezentuje wybór akcji a maksymalizującej wartość.

Techniki wykorzystujące równania Bellmana

**Iteracja wartości:**

Rekurencyjnie oblicza V(s) dla wszystkich stanów, aż do zbieżności.

Po zakończeniu procesu wyznacza optymalną politykę π∗(s).

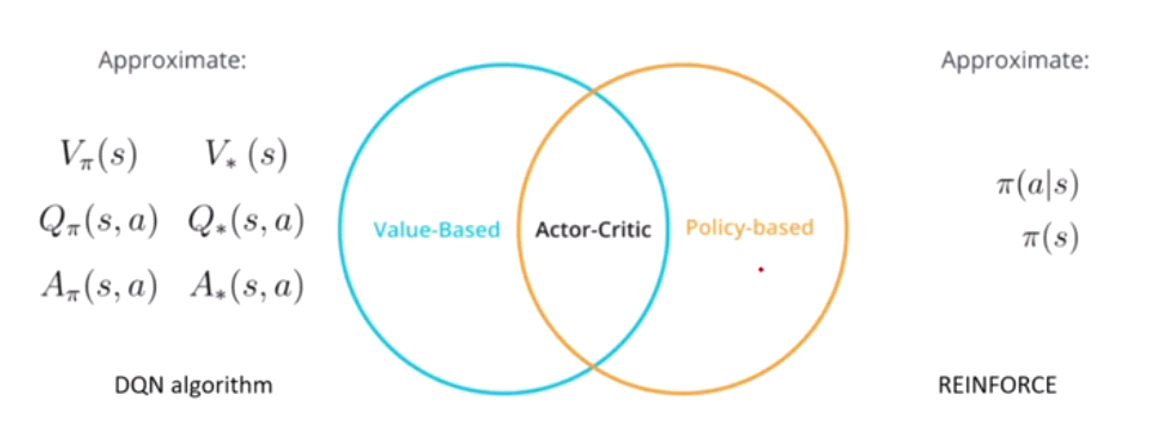
**Iteracja polityki:**

Naprzemienne kroki oceny polityki (Vπ(s)) i jej ulepszania (π′(s)).

Równania Bellmana są używane w obu etapach.

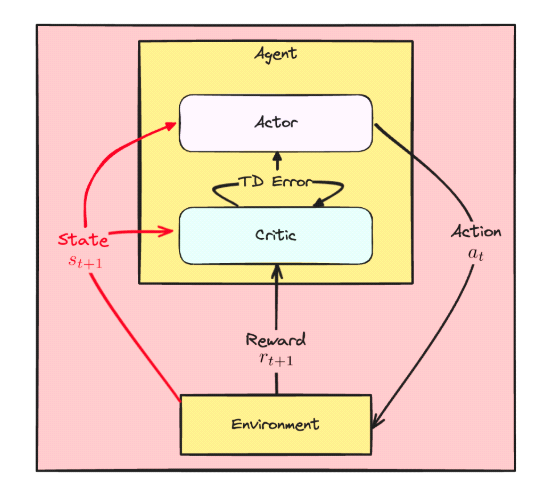
Równania Bellmana są podstawą algorytmów uczenia przez wzmacnianie, ponieważ umożliwiają propagację informacji o nagrodach w czasie i ocenę długoterminowych konsekwencji działań agenta

**5.8 Algorytm Aktor-Krytyk (Actor-Critic)**



*Rysunek 4 A brief review of Actor Critic Methods,* [*https://www.youtube.com/watch?v=aODdNpihRwM*](https://www.youtube.com/watch?v=aODdNpihRwM)

**Algorytm aktor-krytyk** jest połączeniem algorytmów aproksymacji funkcji polityki (policy function) i funkcji wartości (value function) (Rysunek 4). W algorytmach opartych na polityce typu REINFORCE, funkcja polityki jest aktualizowana na końcu epizodu, co jest mało efektywne. Wysoka wariancja gradientu (rezultat sumowania wszystkich zdarzeń z epizodu) powoduje, że potrzeba więcej próbek (epizodów) celem stabilizacji modelu.  
 Algorytm aktor-krytyk rozwiązuje ten problem, korzystając z **metody różnicy czasowej** (ang. *Temporal Difference*). Dzięki temu uczy się przy każdym kroku, a nie tylko na końcu epizodu. (rysunek 5 przedstawia dynamikę procesu)



Pomysł polega na wprowadzeniu agenta zbudowanego z dwóch elementów:

**Aktora** - uczy się polityki π(a∣s), która określa, jakie akcje należy podejmować w danych stanach.

**Krytyka** - Szacuje wartość stanu V(s) i ocenia, jak dobra była decyzja aktora.

**Różnica czasowa** - Krytyk oblicza błąd różnicy czasowej δt ​, który służy jako sygnał wzmocnienia do ulepszania polityki w aktorze.

*, (wzór 8)*

*gdzie:*

*δt ​ to błąd różnicy czasowej (TD-error),*

*rt​ to nagroda natychmiastowa,*

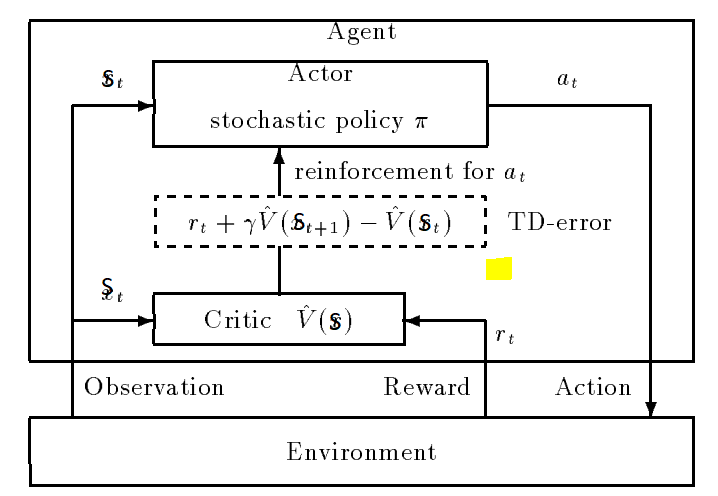
*V(s) to funkcja wartości stanu,*

*γ to współczynnik dyskontowania.*

Metody aktor-krytyk łączą zalety:

Metod opartych na wartościach (redukcja wariancji dzięki krytykowi),

Metod opartych na politykach (elastyczność w modelowaniu przestrzeni ciągłych).

Na rysunku 7 widzimy dokładniej przebieg algorytmu aktor-krytyk

*Rysunek 6: Ha jime Kimura, Shigenobu Kobayashi An Analysis of Actor/Critic Algorithms using Eligibility Traces: Reinforcement Learning with Imp erfect Value Functions:* [*http://users.umiacs.umd.edu/~hal/courses/2016F\_RL/Kimura98.pdf*](http://users.umiacs.umd.edu/~hal/courses/2016F_RL/Kimura98.pdf)

Opis formalny algorytmu uwzględniającego wykorzystanie sieci neuronowych zaczerpnięty z „Reinforcement Learning: An Introduction”

**Wejście:**

π(a∣s,θ), różniczkowalna funkcja prawdopodobieństwa wyboru akcji a w stanie s.

V(s,w), różniczkowalna funkcja szacująca wartość stanu s.

Współczynniki uczenia: αθ>0, αw>0.

**Inicjalizacja:**

Parametry polityki: θ∈R.

Wagi funkcji wartości: w∈R.

**ALGORYTM:**

**Pętla nieskończona (dla każdego epizodu):**

1. Inicjalizuj **s** pierwszy stan epizodu.

2. I←1

**Pętla czasowa (dopóki sss nie jest terminalny):**

3. Wybierz akcję a∼π(⋅∣s,θ).  
4. Wykonaj akcję a, zaobserwuj nowy stan s′ i nagrodę r.  
5. Oblicz błąd TD (δ):

(nawiązanie do wzoru 8)

\*(Jeśli st+1 jest stanem terminalnym, to V(s′,w)=0.

6. Zaktualizuj wagi funkcji wartości:

7. Zaktualizuj parametry polityki:

8. Zaktualizuj współczynnik wpływu I:

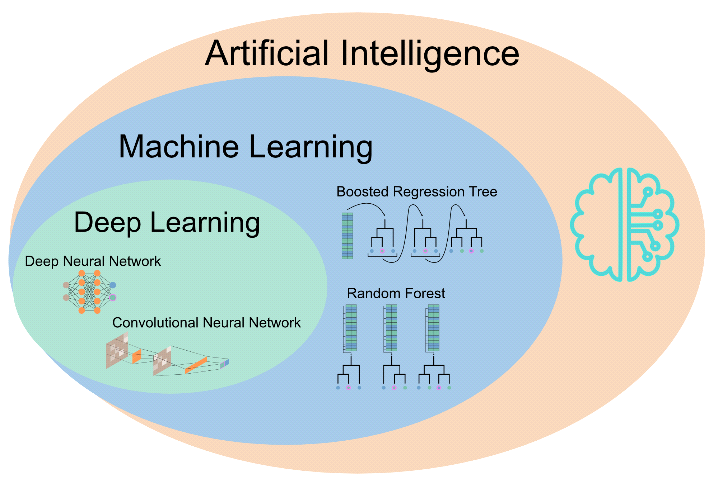
I←γI

9. Przejdź do następnego stanu:

s←st+1

**5.10 Deep Learning w kontekście RL**

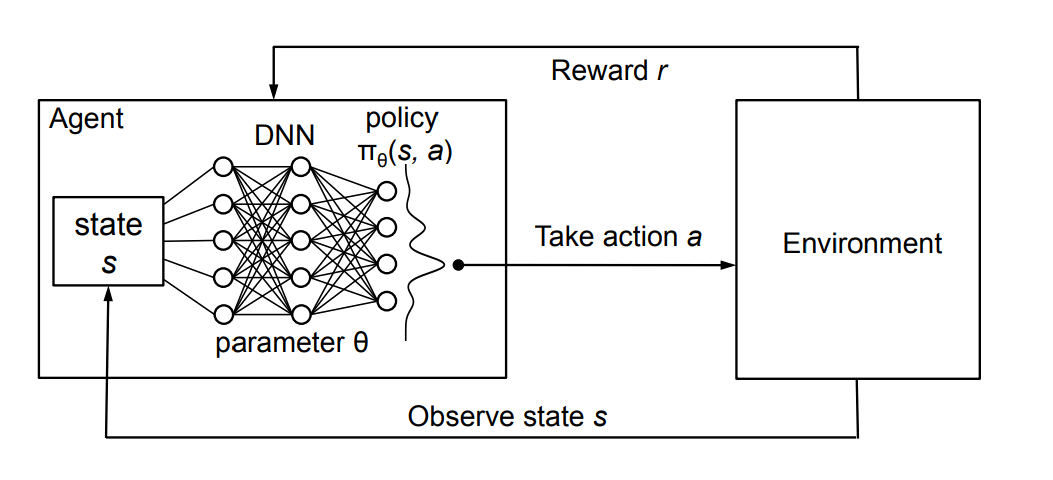
**Uczenie głębokie (Deep Learning, DL)** to dziedzina sztucznej inteligencji, która korzysta z wielowarstwowych sieci neuronowych (rysunek 3), pozwalających na efektywne przetwarzanie i predykcje złożonych funkcji. W uczeniu przez wzmacnianie, metody DL odgrywają kluczową rolę w rozwiązywaniu problemów związanych z dużymi i złożonymi przestrzeniami stanów i akcji.Klasyczne metody, wyznaczanie polityki lub wartości, polegają na iteracyjnym wykonywaniu równań Bellmana (wzór 6,7) w celu propagacji nagród w czasie. Dzięki wykorzystaniu sieci neuronowych, takie obliczenia mogą zostać „nauczone”, co redukuje koszt obliczeniowy do jednorazowego wytrenowania modelu.



*Rysunek 8* [*Maximilian Pichler and Florian Hartig*](https://www.uni-regensburg.de/biologie-vorklinische-medizin/theoretische-oekologie/mitarbeiter/maximilian-pichler/)*, Machine Learning and Deep Learning with R,* [*Maximilian Pichler and Florian Hartig*](https://www.uni-regensburg.de/biologie-vorklinische-medizin/theoretische-oekologie/mitarbeiter/maximilian-pichler/)*,* [*https://theoreticalecology.github.io/machinelearning/*](https://theoreticalecology.github.io/machinelearning/)

W 2015 Firma irma Google DeepMind zaprezentowała, jak głębokie konwolucyjne sieci neuronowe (Convolutional Neural Network) mogą automatyzować ekstrakcję cech, umożliwiając RL radzenie sobie z zadaniami wymagającymi rozumienia zdarzeń w przestrzeni. Przełomem okazało się opracowanie sieci Deep Q-Network (DQN), która łączyła Q-learning z głęboką CNN. Architektura ta pozwoliła DQN na uczenie się wartości Q(s,a) bezpośrednio z surowych danych wejściowych, takich jak piksele. DQN udowodniła swoje możliwości, ucząc się grać w 49 różnych gier Atari i osiągając lub przewyższając poziom człowieka w wielu z nich.

Na rysunku (9) pokazano ogólny schemat zastosowania sieci neuronowej (DNN) do predykcji polityki πθ, (*algorytm Policy Gradient***)** gdzie agent wchodzi w interakcję ze środowiskiem. Należy zwrócić uwagę na symbol θ będący parametrem sieci.



*Rysunek 9 Reinforcement Learning with policy represented via DNN, Hongzi Mao, Mohammad Alizadeh, Ishai Menache, Srikanth Kandula;* [*https://people.csail.mit.edu/hongzi/content/publications/DeepRM-HotNets16.pdf*](https://people.csail.mit.edu/hongzi/content/publications/DeepRM-HotNets16.pdf)

**6. Pakiet SUMO**

Eclipse SUMO to darmowy, otwartoźródłowy pakiet do modelowania systemów transportu intermodalnego, w tym pojazdów drogowych, transportu publicznego oraz ruchu pieszych. Projekt został zainicjowany w 2001 roku przez pracowników Instytutu Systemów Transportowych Niemieckiego Centrum Lotnictwa i Kosmonautyki (DLR).

SUMO jest zestawem aplikacji oferując narzędzia do generowania i importowania sieci drogowych z różnych formatów, a także do tworzenia scenariuszy o dużej skali, takich jak symulacje ruchu w miastach. Symulacje w SUMO są mikroskalowe co oznacza, że każdy pojazd jest modelowany osobno, ma swoją własną trasę i porusza się indywidualnie. Scenariuszach mają możliwość wprowadzana losowości zdarzeń.

SUMO znajduje zastosowanie w badaniach nad komunikacją V2X (pojazd-pojazd i pojazd-infrastruktura). Generowane scenariusze służą do oceniania algorytmów wyboru tras, dynamicznej nawigacji i optymalizacji sygnalizacji świetlnej.

Platforma posiada modele emisji hałasu oraz zanieczyszczeń powietrza, umożliwiając ocenę ekologicznych skutków transportu. Obsługuje również wsparcie dla pojazdów autonomicznych.

Do komunikacji z SUMO w czasie rzeczywistym najczęściej wykorzystuje się interfejs TraCI (Traffic Control Interface), działający jako usługo TCP/IP. TraCI umożliwiający odczytywanie parametrów symulacji oraz inicjowanie zmieniających się parametrów środowiska.

SUMO jest popularny dzięki wszechstronności, otwartemu kodowi źródłowemu oraz wsparciu dla dużych symulacji. Dzięki API platformę można integrować z innymi narzędziami poprzez biblioteki w języku Python, C++, JavaMATLAB