УДК 004.896
DOI https://doi.org/10.32782/2663-5941/2024.1.1/16

Бойчук В.О.
Хмельницька загальноосвітня школа № 14

Бойчук А.А.
ПВНЗ «Хмельницький економічний університет»

Бойчук М.В.
EPAM Ukraine

ПРО ФОРМУВАННЯ ПОСЛІДОВНОСТЕЙ ДІЙ У ІНТЕЛЕКТУАЛЬНИХ АГЕНТІВ

Сучасні роботи та інші інтелектуальні агенти досить часто наражаються на проблеми з організацією автономного існування в реальних середовищах. У той же час процеси організації і планування дій, які первісно притаманні як людям, так і доволі простим біологічним істотам, дозволяють їм там виживати. Тому у статті описуються уявлення психології та нейрофізіології про організацію процесів формування і планування послідовностей дій у людей і тварин. Процеси формування і планування послідовностей дій розглядаються на основі декількох рівнів рефлекторних кілець, що дозволяє гнучко формувати послідовності дій.

Показано, як може реалізуватися дана специфіка на базі систем інтелектуального планування, мовної моделі Transformer, глибоких штучних нейронних мереж, спайкових нейронних мереж, навчання з підкріпленням, моделей штучних емоцій.

Враховуючи недоліки вищевикладених технологій на базі біонічного підходу з використанням моделі на основі штучних емоцій представлено метод для формування послідовностей дій у інтелектуальних агентів. Агент в кожний момент часу виконує і модифікує параметри політики вибору нової дії на основі значень виходів класифікаторів і вектору значень станів.

З використанням представленого методу показано порядок формування ланцюжків безумовних рефлексів. Визначено етапи формування умовних рефлексів на основі безумовних рефлексів. Функціонування орієнтовного безумовного та умовного рефлексів продемонстровано для моделей роботів в середовищі V-REP.

Метод розроблений на основі аналогій з функціонуванням мозку для гнучкого багаторівневого формування послідовностей дій у інтелектуальних агентів і надасть змогу забезпечити їх функціонування в складному змінному середовищі.

Ключові слова: *штучний інтелект, нейронна мережа, інтелектуальний агент, планування, штучна емоція, навчання з підкріпленням, рефлекс.*

Постановка проблеми. Проблемі вивчення організації і планування дій присвячено багато праць у різних сферах науки: математиці, інформатиці, штучному інтелекті (ШІ). І не зважаючи на це сучасні роботи та інші інтелектуальні агенти наражаються на проблеми з організацією автономного існування в реальних середовищах. У той же час процеси організації і планування дій, які первісно притаманні як людям, так і доволі простим біологічним істотам, дозволяють їм там виживати.

Психологія і нейрофізіологія розглядають організацію цих процесів доволі всебічно і комплексно, хоча часто і в узагальненій описовій формі. Наприклад у [1] організація і планування дій людини розглядається на основі декількох рівнів рефлекторних кілець. На найнижчому рівні забезпечується несвідоме регулювання фізіологічних параметрів, наприклад тонусу м'язів з організацією їх тремору при переохолодженні. На більш високому рівні стоять найпростіші безумовні рефлекси, наприклад колінний, де рефлекторна дуга формується тільки у спинному мозку. Далі йдуть стандартні безумовні переміщувальні рефлекси, такі як хода. Вони більш складні і в їх організації приймають участь більш високі відділи мозку. Вище стоять сформовані під керуванням кори мозку нестандартні послідовності дій. Вони можуть мінятися і перегруповуватись під конкретну ціль. На найвищому рівні мова

і мислення, які характеризуються абстрактним і узагальненим змістом по плануванню і організації дій. Тобто з кожним рівнем йде ускладнення дій і цілей, які за допомогою цих дій досягаються, від регулювання фізіологічних параметрів організму до планування і реалізації мети діяльності організму на деякий період. У організації конкретної дії як правило беруть участь декілька рівнів рефлекторних кілець.

Людина може свідомо оперувати діями за безпосередньої участі процесів мислення і мови. При повтореннях послідовності дій з успішним завершенням йде оволодіння нею і вона може переміщуватись на нижчий неусвідомлений рівень з формуванням динамічного стереотипу (складного умовного рефлексу).

В праці [2] наведена схема такого ланцюжкового умовного рефлексу, де дії в ланцюжку при навчанні стимулюються комбінацією екстероцептивних та пропріоцептивних ефекторів з утворенням умовних комбінаційних центрів на базі виконання підтвердження. Наведені експериментальні дані, які показують, що у черепах виробляються такі рефлекси тільки максимум з трьох ланок і з 116 повторів, у людини з 5 ланок з одного повтору.

Така багаторівнева схема з формуванням динамічних стереотипів дає змогу планувати і керувати діяльністю систем з багатьма ступенями свободи в складному середовищі, яке постійно змінюється. У випадку коли виникають труднощі у користуванні сформованим стереотипом відбувається його деавтоматизація, перехід з неусвідомленого рівня на усвідомлений.

Аналіз останніх досліджень і публікацій. Сучасні засоби ШІ не реалізують цю схему повністю. Розглянемо їх особливості і можливості по її реалізації починаючи з мовного рівня.

Системи інтелектуального планування (GPS, Strips, Warplan, Interplan, Tweak) використовують мовний рівень, більшість у формалізованому вигляді математичної логіки[3]. Формування планів зводиться до логічного висновку. Така формалізація класичного планування без зв'язку з нижчими рівням дає змогу цим системам працювати тільки в добре структурованих областях за допомогою експертів. Але у класичному плануванні є багато елементів, які використовує і людина, оперування підцілями, відкати від помилкових дій, пробне виконання поточного плану, прямий та зворотний пошук в просторі станів та ін.

Сьогоднішні системи на основі популярної мовної моделі Transformer [4] мають досить потужні можливості по генерації текстів по заданій тематиці і візуальних зображеннях, створенню вербальних планів дій по завданню користувача.

З прикладів такого використання в плануванні можна назвати архітектури Decision Transformer і Trajectory Transformer у яких оминають звичайний для навчанням з підкріпленням процес максимізації віддачі та безпосередньо генерують ряд майбутніх дій, які забезпечують бажаний прибуток [5, 6]. Також модель Transformer використовується для аналізу двовимірних карт перешкод для отримання шляхів руху роботів [7,8].

Слід вказати схожі сучасні розробки [9, 10, 11], один з яких проект Google DeepMind [12], де на основі багатьох експериментів та інформації з Інтернету формується візуально-мовна модель (VLM). На її основі в проекті Google DeepMind розроблена модель RT-2 (Robotic Transformer), яка дозволяє генерувати послідовність дій робота по виконанню елементарних для людини дій, на кшталт викидання сміття у корзину. Навчанням звичайно проводиться на основі демонстрації цих дій з надаванням роботу текстового опису дій і вихідних значень класифікаторів фото і відео цих демонстрацій з комбінуванням їх у моделі RT-2. При виконанні на кожному кроці обчислюються значення Q-функцій і відповідно них вибираються параметри дій. Використання мовної моделі Transformer дає змогу виконувати узагальнення і розмірковування, наприклад сміттям може бути не тільки один об'єкт, відслідковувати невидимі об'єкти і т.д. Дана розробка є вагомим кроком у спробах добитись функціонування агентів у реальних середовищах. Дії в ній безпосередньо прив’язані до мовної моделі Transformer, яка має потужні можливості по «розмірковуванню». Але таким чином пропускаються необхідні проміжні рівні з описаних на початку статті, які забезпечують гнучкість і ефективність навчання. Крім того в більшості випадків біологічні організми вчаться самостійно і їм навіть непотрібно для цього використання мови.

Якщо перейти на рівень кори головного мозку, яка згідно психології і нейрофізіології грає важливу роль у плануванні нових дій, то її моделлю в сучасному ШІ є глибокі штучні нейронні мережі, переважною функцією яких є задача класифікації. Вони широко і успішно використовуються у різних областях. Однак високий рівень абстракції штучних нейронів порівняно з біологічним аналогом разом із відсутністю в них можливості відобразити часову динаміку біологічних нейронів не дає змогу тільки ними описати послідовності дій і їх формування. Тому для опису планування

послідовності дій їх треба доповнювати іншими моделями.

Завдяки здатності відобразити різноманітну динаміку біологічних нейронів, представити час, частоту, фазу, спайкові нейронні мережі потенційно здатні моделювати складні послідовності процесів обробки інформації, що відбуваються в мозку. Однак теоретичний апарат спайкових нейронних мереж зараз знаходяться в стадії становлення. Із використанням спайкових нейронних мереж моделюються безумовні і умовні рефлекси. В [13] за допомогою спайкових нейронних мереж проводиться детальна структурна і функціональна імітація мозочка, що дає змогу інтелектуальному агенту наприклад вчитися переводити погляд точно на об'єкт у полі зору. Але моделювання ведеться на досить низькому рівні. Опис складних послідовностей дій за допомогою цих моделей поки що занадто громіздкий.

При пошуку послідовності дій до цілі також використовується навчання з підкріпленням – один із розділів машинного навчання в ході якого випробувана система навчається, взаємодіючи з деяким середовищем [14]. Цей метод моделює життя біологічної істоти, яка починаючи з народження та протягом усього життя використовує саме механізм навчання з підкріпленням. У ссавців цю роботу виконує дофамінова система. Її робота не до кінця вивчена, але зводиться до того, що у разі отримання нагороди, дофамінова система через механізми пам'яті закріплює зв'язки між нейронами, які були активні безпосередньо до цього. Навчання з підкріпленням зазвичай описується у формі марківського процесу прийняття рішень. Агент отримує відкладену винагороду на наступному часовому кроці, щоб оцінити свою попередню дію. Доцільність кожного напрямку руху в просторі станів рахується за рівнянням Белмана. Методи навчання з підкріпленням Q-learning, SARSA, DQN, DDPG часто використовуються в моделюванні ігор, з досить високою продуктивністю на рівні або навіть вище людини. Однак навіть найкращі алгоритми навчання з підкріпленням вимагають десятки мільйонів кроків для навчання на просторі станів, які можна порівняти за ефективністю з випадковим пошуком, відповідно це дає змогу працювати лише на порівняно низьких розмірностях. Тому останнім часом ці алгоритми використовуються у комбінації з нейронними мережами.

Зараз на фоні бурхливого розвитку штучних нейронних мереж порівняно мало уваги приділяється механізму емоцій, хоча в природі вони відповідають за навчання з підкріпленням. Досить повний огляд праць, де використовують моделі емоції, можна знайти в статті [15]. Але представлені методи ще далекі від широкого практичного використання і в дечому відірвані від попередніх підходів.

Формулювання цілей статті. Тому враховуючи недоліки вищевикладених методів потрібна біонічна модель, яка дозволяла б описати динаміку, на відміну від штучних нейронних мереж. І яка б описувала складні багатоланцюжкові послідовності дій на декількох рівнях, на відміну від спайкових нейронних мереж. І могла б використовувати навчання з підкріпленням на основі емоцій по аналогії з функціонуванням біологічних організмів.

Виклад основного матеріалу. Стаття продовжує і розширює опис методу викладеного у [16, 17, 18], де зокрема представлено модель послі-довності дій у вигляді направленого графа $G=(V,E)$ у якому кожній вершині v відповідає виконувана дія a з множини елементарних дій A, а кожній на-правленій дузі e, по якій відбувається перехід між діями, вагова функція $w:E\rightarrow R$, яка відображає ребра на їх ваги.

Вага дуги залежить від значень станів агенту, які є аналогами емоцій і почуттів та в загальному відображають функції лімбічної системи людини і вплив різних нейромедіаторів на відділи мозку згідно шляхів їх розповсюдження для визначення загальної важливості для агента зовнішніх або внутрішніх подразників [19]. Стани агенту $Y=(y_1,y_2,\ldots,yn),\ 0\leq y_i\leq 1$ можуть змінюватись класифікуючими елементами, які отримують інформацію як з зовнішнього середовища, так і на основі внутрішніх характеристик агента. Кожен стан y_i після зміни характеризується динамікою у часі. На початку діяльності агент може змінювати стани стандартним чином на визначений перелік подразників. В процесі діяльності список таких подразників і реакції можуть змінюватись.

Класифікуючі елементи агента є множиною штучних нейронних мереж з виходами $X_1, X_2,\ldots,X_m$, *де* $X=(x_1,x_2,\ldots,x_n),\ 0\leq x_i\leq 1$.

Значення X можуть бути:

1) бінарними виходами навченої штучної нейронної мережі. Цей варіант ймовірно відображає функціонування в реальних біологічних системах;

2) точкою в схованому багатовимірному просторі на виході штучної нейронної мережі з зафіксованими вагами. Даний варіант використовується в наступному моделюванні.

Формування нових послідовностей дій відбувається за рахунок налаштування коефіцієнтів,

які відображають важливість кожного стану для дуги. Цей процес виконується циклічно і як правило починається при підвищених рівнях негативних станів та закінчується досягненням позитивних станів. Після цього послідовності можуть бути використовуватись в функціонуванні агенту при переважанні нормального стану, рівень якого моделює вплив нейромедіатора серотоніну. При неуспіху послідовності агент впадає в стан невдачі і вона може поступово розформуватися за рахунок переналаштування коефіцієнтів та зміни рівня нормального стану.

В експерименті статті [16] модель робота рухалась по простору оточеному стінами. Значення w_i в експерименті лінійно залежали від станів y_j. Стан «страху» використовувався для самонавчання послідовності дій по уникненню зіткнень зі стінами на основі стандартної дії при цій емоції. Напрямок повороту після навчання вибирався згідно класифікаційних значень, отриманих в процесі навчанні з інфрачервоних давачів дальності при успішних поворотах. Основна увага приділялась алгоритму запам'ятовування послідовності дій.

Дана стаття акцентує увагу на використанні класифікаційних значень X, що дозволить описувати багаторівневу схему описану на початку статті.

Агент в кожний момент часу t виконує і модифікує параметри політики π вибору нової дії a_i на основі вхідних даних $\gamma_t(X_t, Y_t)$ і досягнення станів:

$$\pi(\gamma_t) = \pi(X_t, Y_t) \rightarrow a_i \in \boldsymbol{A},$$

де X_t – значення виходів класифікаторів у момент часу t, Y_t – вектор значень станів у момент часу t.

Тобто вибір наступної дії залежить від класифікаційних значень X_t, отриманих з зовнішніх сенсорів і внутрішніх рецепторів та станів агенту Y_t.

Якщо розглядати виконання даної політики на графовій моделі, то маємо вагові функції ребер $w{:}E{\rightarrow}R$ на орієнтованому графі G. Нехай існує вершина v_i і множина ребер E_i^-, що виходять з вершини. Тоді шлях згідно політики π утворюється ітераційно згідно вибору:

$$w(e_i)\, e_i E_i^- w(e_i) > w_{thres} v_{i+1},$$

де w_{thres} – порогове значення ваги.

Для реалізації даної політики використаємо поняття осередок збудження і введемо поняття відповідності. В якості осередків збудження розглядались виходи класифікаторів моделі, яким відповідають нейрони детектори кори головного мозку, та елементарні дії, яким відповідають функціонуючі мотонейрони і м'язи. Між осередками збудження встановлюються зв'язки при близьких часових інтервалах активації на основі моделювання нейронної пластичності. Відповідності R встановлюються між деякою дугою e_i і значенням x виходу класифікатора і забезпечують виконання ланцюжків рефлексів.

Відповідності можуть реалізовуватись:

1) встановлюватись при початковому налаштуванні агента і можуть активізуватись або у початковий момент часу або у через якийсь час згідно деякої умови;

2) при навчанні агента на деякий час.

На рис. 1 умовно прямокутниками показані зображення, що їх може сприймати агент і виходи класифікатора x_1, x_2, x_3, x_4, x_5. Їх відповідності r_1, r_2, r_3, r_4, r_5 з дугами з вагами w_1, w_2, w_3, w_6, w_7 показані двонаправленими стрілками.

Відповідність R встановлюється до конкретного стану на дузі і підвищує ймовірність вибору дуги для переходу до наступної дії на основі виконання умови:

$$\forall\, w_j\, \exists\, y_i > y_{thres} \mid \exists\, x_j\, R\, y_i \Rightarrow y_i = y_i + \Delta_{thres},$$

де y_{thres} – значення стану при якому починає працювати відповідність,

Δ_{thres} – значення яке додається до стану.

В безумовних рефлексах відповідні значення виходів класифікаторів постійно прив'язані до визначених дуг і забезпечують перехід по цим дугах сумісно з відповідними станами агенту. Основну роль у запуску рефлексів має встановлення рівня деякого стану y_i з комбінації станів $Y_{US}0$ більше деякого порогу. Після цього очікується надходження потрібних значень на виходах класифікаторів $X_{US}0$ для запуску рефлексу. Встановлення такого стану може викликатися внутрішніми і зовнішніми факторами, наприклад виділенням гормонів андрогенів у людей або падінням заряду батареї у агента.

Умовно формування ланцюжку безумовного рефлексу можна показати таким чином:

$$(Y_{US0} \rightarrow X_{US0}) \rightarrow ((X_{USI}, Y_{USI}) \rightarrow a_1,$$
$$(X_{US2}, Y_{US2}) \rightarrow a_2, \ldots, (X_{USn}, Y_{USn}) \rightarrow a_n).$$

При цьому формується шлях з послідовності дій і ваг дуг: $S_{US} = (a_1, w_1, a_2, w_2, \ldots, w_{n-1}, a_n)$.

Умовні рефлекси утворюються на основі безумовних або інших умовних рефлексів. В [2] описані такі способи утворення ланцюжків умовних рефлексів у тварин:

– об'єднання в ланцюг екстероцептивних одиночних подразників одиночних рухових реакцій;

– нарощування ланцюга рухів з кінця;

– вклинювання нових рухів і подразників подібним чином, але між останнім ланкою ланцюга і підкріпленням;

– при формуванні ланцюга рухів тварин не обмежують у рухах, але підкріплюють ті ланцюги рухів, які були «правильними».

Початок формування умовних рефлексів задається підвищеними рівнями негативних станів і виконується за рахунок зміни станів та значень класифікаторів по ланцюжку дій у S_{CS}, який утворюється.

Якщо формування йде на основі безумовного рефлексу то заміну значень виходів класифікаторів на нові умовні можна показати так:

$$(X_{CS0}) \rightarrow (X_{US1} \rightarrow X_{CS1}) \rightarrow (X_{US2} \rightarrow X_{CS2}) \rightarrow \ldots \rightarrow (X_{USn} \rightarrow X_{CSn}).$$

При формуванні умовного рефлексу процес встановлення відповідності між деякою дугою w і вихіднім значенням класифікатора x відбувається при досягненні агентом позитивних станів після виконання послідовності дій по шляху S_{CS}.

При цьому для встановлення відповідності на кожному циклі навчання на змінному виході класифікатора X має виконуватись умова:

$$\forall\, w_j \in S_{CS}\ \exists\ U(X, \varepsilon),$$

де ε – окіл точки X.

Процес встановлення відповідності розповсюджується при навчанні по ланцюжку активованих дій з формуванням значень Δ_{thres} сумісно з коректуванням коефіцієнтів k [17, 18], які відображають вагу кожного стану для дуги.

Умовно виконання ланцюжку сформованого умовного рефлексу можна показати так:

$$(X_{CS0} \rightarrow Y_{CS0}) \rightarrow ((X_{CS1}, Y_{CS2}) \rightarrow a_1,$$
$$(X_{CS2}, Y_{CS3}) \rightarrow a_2, \ldots, (X_{CSn}, Y_{CSn}) \rightarrow a_n)$$

Однак послідовність дій не є повністю автоматичною, якщо значення X_{CS} отримуються з зовнішніх сенсорів. Агент при її виконані в кожному циклі моделювання повинен очікувати відповідних значень класифікаторів. Назвемо цю послідовність дій напівавтоматичною.

Наступний можливий етап навчання – забезпечення формування автоматичної послідовності. При цьому умовні значення X_{CS}, які отримуються з зовнішнього середовища, заміняються значеннями X_{IS}, які генеруються на основі показників внутрішніх рецепторів агенту. У живих організмах приблизно таку функцію при автоматизації дій виконує мозочок.

Якщо для дуг шляху S_{CS} в кожному циклі навчання існують значення X_{CSi} з зовнішніх сенсорів і XISj з внутрішніх рецепторів, що:

$$\forall w_j \in S_{CS}\ \exists X_{CSi} \in U(X_{CSi}, \varepsilon_i) \mid \exists X_{ISj} \in U(X_{ISj}, \varepsilon_j) \Rightarrow$$
$$X_{CSi} \rightarrow X_{ISj}$$

Заміну значень виходів класифікаторів на основі внутрішніх рецепторів можна показати так:

$$(X_{CS1} \rightarrow X_{IS1}) \rightarrow (X_{CS2} \rightarrow X_{IS2}) \rightarrow \ldots \rightarrow (X_{CSn} \rightarrow X_{ISn}).$$

А сам сформований умовний рефлекс тепер буде виглядати так:

$$(X_{CS0} \rightarrow Y_{CS0}) \rightarrow ((X_{IS1}, Y_{IS1}) \rightarrow a_1, (X_{IS2}, Y_{IS2}) \rightarrow a_2, \ldots,$$
$$(X_{ISn}, Y_{ISn}) \rightarrow a_n).$$

Після встановлення відповідності значення на виходах класифікаторів X_{CS0} разом з відповідним рівнем стану Y_{CS0} використовується для ідентифікації існування ланцюжка дій і початку виконання послідовності. Перехід підтримується надалі при цільовому завершенню послідовності дій. При нецільовому завершенню агент попадає у стан «невдачі», що дає змогу коректувати даний або шукати інші ланцюжки. Будемо називати пару (X_{CS0}, Y_{CS0}) точкою входу в рефлекс.

При повному виконанні навчання умовному рефлексу при звертанні до нього за допомогою точки входу (X_{CS0}, Y_{CS0}) в комбінації станів Y_{CS0} переважає нормальний стан. При незавершеному навчанні X_{CS0} може відповідати комбінація Y_{CS0} з переважанням негативних станів.

Розглянемо можливий варіант реалізації безумовного орієнтовного рефлексу (рис. 1) і формування на його основі умовного рефлексу.

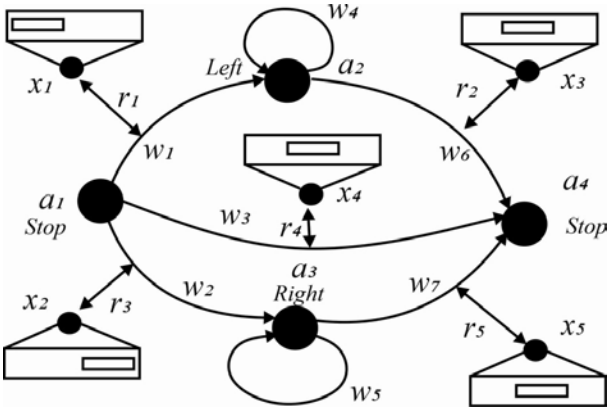

Рис. 1. Граф реалізації безумовного орієнтовного рефлексу

Біологічний орієнтовний рефлекс полягає в тому, що тварина або людина повертається до нового незвичного об'єкта, щоб він опинився у центрі поля зору. Згідно нейрофізіології на однаковий об'єкт даний рефлекс гальмується через 10–15 повторів.

Виконання агентом орієнтовного рефлексу починається з ідентифікації аномалії, незвичного об'єкту в полі зору агенту. При цьому зразу збільшується значення стану «здивування», що відповідає генерації осередку збудження у мозку. Спочатку згідно рефлексу виконується так звана стоп-реакція. Після цього новий об'єкт переводиться в центр поля зору агенту. Для цього вико-

нуються повороти в необхідну сторону в залежності від поточного положення нового об'єкта і відповідно від класифікаційних значень x_1, x_2, які стимулюються рівнем поточного стану «здивування». Закінчуються повороти коли об'єкт буде в середині поля зору агенту при відповідному значенні x_3 або x_5. Далі йде зупинка і генералізований орієнтовний рефлекс закінчується.

Дані дії виконуються за рахунок відповідного початкового налаштування коефіцієнтів k у дугах моделі при відповідному стані агенту і завчасного встановлення відповідності з значеннями X.

Функціонування орієнтовного рефлексу було промодельовано з використанням стандартного робота DR20 з середовища V-REP EDU з відеокамерою, ультразвуковим та інфрачервоними давачами дальності. Робот має два коліщатка, які керуються за допомогою двигунів (рис. 2).

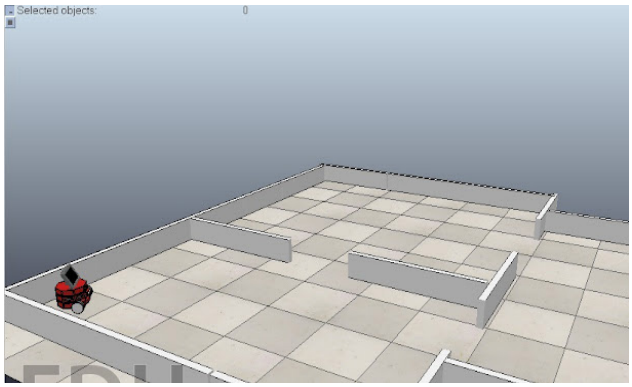

Рис. 2. Вигляд робота DR20 і середовища моделювання рефлексів

Для керування роботом використовується так званий зовнішній скрипт на мові Python, який взаємодіє з роботом через мережевий сокет. При цьому застосовувались бібліотеки OpenCV, Tensorflow, Keras, Sklearn. Програма запускалась на настільному комп'ютері з процесором Ryzen 5 3600, 32 Гб загальної пам'яті, відео Palit 1660 6 Гб пам'яті.

Для реалізації рефлексу робот повинен:

1) запам'ятати стандартне середовище;

2) ідентифікувати появу нового об'єкта;

3) повернутися на нього.

Камерою робота при русі з уникненням перешкод отримуємо набір зображень оточуючих стін. На основі зображень формується масив, що характеризує стандартне середовище робота, як виходи замороженої нейронної мережі ResNet-50. Використовуючи бібліотеку Sklearn.cluster на основі отриманого масиву згідно з алгоритму кластеризації DBSCAN формуються кластери, що характеризують стандартне середовище робота.

Коли в стіні з'явиться отвір (новий об'єкт) робот виявляє аномалію за допомогою функції dbscan_predict, яка переводить його в емоційний стан «здивування» і він переходить до виконання дій згідно рис. 1. Щоб визначити розташування нового об'єкта та повернутись до нього, робот використовує попередньо навчену згорткову нейронну мережу для трьох положень об'єкта: ліворуч, праворуч і по центру. При повороті в кожному циклі моделювання контролюється момент, коли об'єкт опиниться в центрі поля зору

Розглянемо вироблення складного умовного рефлексу на основі даного безумовного рефлексу, коли робот навчається проїжджати через отвір у стіні.

Вибираємо кількість повторень навчання n=10, по кількості необхідній для гальмування орієнтовного рефлексу на однаковий об'єкт. Відправляємо робота переміщатися приблизно під одним кутом до стіни і в його поле зору попадає новий об'єкт (отвір у стіні), який і викликає у нього орієнтовний рефлекс, описаний раніше (рис. 3).

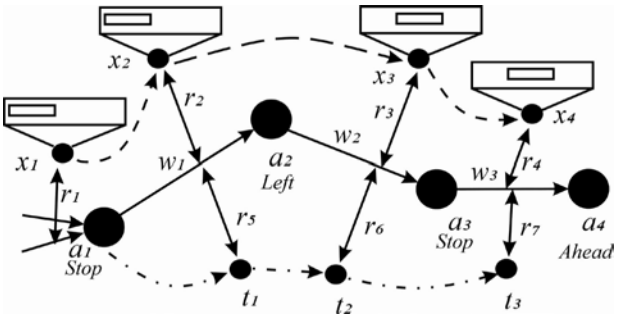

Рис. 3. Граф формування умовного рефлексу на основі безумовного орієнтовного рефлексу

Виконується зупинка a_1, поворот ліворуч a_2 на отвір, зупинка a_3 і далі генералізований орієнтовний рефлекс переходить в локальний, що стимулює агента рухатись далі прямо a_4. Після руху прямо робот проїжджає через отвір і переходить в позитивний стан успіху від виникнення нового простору для руху.

По аналогії з біологічним виробленням рефлексу після повторів емоція «здивування» разом з орієнтовним рефлексом на поточне положення отвору гальмується. Запам'ятовується послідовність дій a_1, a_2, a_3, a_4 з оновленням відповідностей дуг w_1, w_2, w_4 поточними значеннями x_2, x_3, x_4 для забезпечення переходів по цим дугам. Також запам'ятовується умова початку виконання рефлексу x_1.

При цьому аналогічно мозку біологічних істот при близьких інтервалах збудження виходів класифікаторів утворюється додатковий асоціативний шлях x_1, x_2, x_3, x_4 між виходами класифікаторів (нейронами-детекторами) по принципу їх акти-

вації в близькі інтервали часу (сусідніх циклах моделювання).

Робот при виконані послідовності в кожному циклі моделювання повинен очікувати відповідного положення отвору, наприклад до досягнення його центра поля зору. Класифікатори (зір робота) зайняті виконанням поточної дії. В процесі навчання умовні значення X_{CS} заміняються часовими значеннями X_{IS}, які генерується на основі вимірювання часових інтервалів внутрішнім годинником агента. Тепер через деяку кількість повторів послідовність дій a_1, a_2, a_3, a_4 стає автоматичною і значення x_2, x_3, x_4 не використовуються для виконання рефлексу. При цьому допоміжну роль у його виконані грає вже ланцюжок t_1, t_2, t_3.

Демонстрація функціонування даних прикладів наведена за посиланням [20].

Тобто при створені умовного рефлексу утворюються три паралельних ланцюжки:

$a_1, a_2, \ldots, a_n$ – власне сформована послідовність дій;

$x_1, x_2, \ldots, x_m$ – ланцюжок виходів класифікаторів зовнішніх сенсорів;

$t_1, t_2, \ldots, t_k$ – ланцюжок виходів класифікаторів часових інтервалів.

Між цими ланцюжками в складному умовному рефлексі утворюються відповідності.

Перший ланцюжок утворюється через навчання на основі емоцій, інші два на його основі по принципу асоціацій при часовій близькості осередків збудження.

Ланцюжок виходів класифікаторів x_1, x_2, x_3, x_4 деякої сформованої автоматичної послідовності дій може не використовуватись при її виконанні. Що надає можливості по його додатковому використанню для прийняття рішень через вплив на стани.

Взагалі формування умовного рефлексу йде аналогічно мозку людини, де спочатку активна права півкуля мозку з «негативними емоціями», а після формування послідовності йде передача виконання до лівої півкулі людського мозку на основі «позитивних емоцій» [21].

В організмі значенням X відповідають класифікуючим нейронам-детекторам кори головного мозку, вершини і дії a – функціонуючим м'язам, дуги і їх ваги w – синапсам нейронів в зв'язках, які утворюються у процесі навчання у мозку, коефіцієнти k – шляхам поширенню медіаторів в мозку і ступеню їх впливу на синапси.

Даний метод дає можливість формувати дії агенту не тільки на основі мови, як сучасних моделях на основі архітектури Transformer, а функціонувати згідно багаторівневої схеми рефлекторних кілець представленої на початку статті.

Висновки. Таким чином у статті уточнені і розширені можливості методу по формуванню послідовностей дій інтелектуальних агентів на основі біонічної моделі з штучними емоціями. Показано як на їх основі можуть бути виконані безумовні і вироблені ланцюжкові умовні рефлекси. Продемонстровано їх функціонування для моделей роботів в середовищі V-REP. Метод і модель розробляються для багаторівневого планування послідовностей дій у інтелектуальних агентів і надає змогу забезпечити їх функціонування в складному змінному середовищі.

В прикладі статті умовний рефлекс вироблявся при зовнішній постановці агента в ідеальні початкові умови з його формуванням на основі безумовного рефлексу з додаванням в його кінці нової дії. При цьому повністю сформувалась автоматична дія. Відповідно у статі не розглядалось функціонування агента при неповністю сформованому ланцюжку дій, переформатування ланцюжків у випадках невдач та механізми використання агентом сформованих умовних рефлексів. Також для розширення можливостей агента треба або збільшувати обсяг навчання, або ввести абстрактні поняття і деяку спрощену мову. Тоді агент сам зможе формувати і оперувати готовими умовними рефлексами. Ці питання будуть темами наступних досліджень і статей.

Список літератури:

1. Wacław Petryński Bernstein's construction of movement model and contemporary motor control and motor learning theories Katowice School of Economics, Katowice, Poland Hum Mov, Vol. 2007; 8(2):136-147.

2. Orienting Reflex and Exploratory Behavior L. G. Voronin, A. N. Leontiev, A. R. Luria, E. N. Sokolov, O. S. Vinogradova, Publisher: American Institute of Biological Sciences, Washington, D.C., 1965. 462 p.

3. Knowledge-level analysis of planning systems A Valente ACM SIGART Bulletin, 1995 dl.acm.org.

4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin Attention is All you Need – 2017. arXiv:1706.03762 [cs.CL].

5. Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivasll, Igor Mordatch Decision Transformer: Reinforcement Learning via Sequence Modeling arXiv:2106.01345 [cs.LG]

6. Michael Janner, Qiyang Li, Sergey Levine Offline Reinforcement Learning as One Big Sequence Modeling Problem 35th Conference on Neural Information Processing Systems (neurips 2021), Sydney, Australia.

7. Devendra Singh Chaplot, Deepak Pathak, Jitendra Malik1 Differentiable Spatial Planning using Transformers, Proceedings of the38th international Conference on Machine Learning, PMLR 139, 2021, arXiv:2112.01010 [cs.LG].

8. Jacob J. Johnson, Uday S. Kalra, Ankit Bhatia1, Linjun Li1, Ahmed H. Qureshi, Michael C. Yip Motion Planning Transformers: A Motion Planning Framework for Mobile Robots arXiv:2106.02791v2 [cs.RO] 13 Nov 2022.

9. Mohit Shridhar, Lucas Manuelli, Dieter Fox CLIPORT: What and Where Pathways for Robotic Manipulation 5th Conference on Robot Learning (CoRL 2021), London, UK. arXiv:2109.12098v1 [cs.RO] 24 Sep 2021.

10. Mohit Shridhar , Lucas Manuelli, Dieter Fox PERCEIVER-ACTOR: A Multi-Task Transformer for Robotic Manipulation 6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand arXiv:2209.05451v2 [cs. RO] 11 Nov 2022

11. RT-1: Robotics transformer for real-world control at scale arXiv:2212.06817v2 [cs.RO] 11 Aug 2023.

12. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control arXiv:2307.15818v1 [cs.RO] 28 Jul 2023.

13. Claudia Casellato, Alberto Antonietti, Jesus A. Garrido, Richard R. Carrillo, Niceto R. Luque, Eduardo Ros, Alessandra Pedrocchi, Egidio D'Angelo Adaptive Robotic Control Driven by a Versatile Spiking PLoS One. 2014; 9(11): e112265. Published online 2014 Nov 12.

14. Aditya Mohan, Amy Zhang. Marius Lindauer Structure in Reinforcement Learning: A Survey and Open Problems arXiv:2306.16021v2 [cs.LG] 9 Aug 2023

15. Emotion in reinforcement learning agents and robots : A Survey Thomas M. Moerland, Joost Broekens , Catholijn M. Jonker. arXiv:1705.05172 [cs.LG] Machine Learning 2017.

16. Бойчук В.О., Бойчук М. В., Жук О. О. Біонічна модель поведінки інтелектуальних агентів Наука й економіка Науково-теоретичний журнал Хмельницького економічного університету Випуск 4 (48), Хмельницький, 2018. С. 137–142.

17. Бойчук В., Бойчук А., Бойчук М., Бурдюг О. Метод формування послідовності дій інтелектуальних агентів. Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка, (66), 2020. С 65–74.

18. Бойчук М. Метод планування послідовності дій на основі навчання з підкріпленням: дипломна робота магістра, ХНУ, -Хмельницький, 2019. 86 с.

19. Ethan S. Bromberg-Martin, Masayuki Matsumoto, Okihide Hikosaka. Dopamine in Motivational Control: Rewarding, Aversive, and Alerting. Neuron. 2010. 68, 815–834.

20. Boychuk Vadym Reinforcement learning robots based on emotions [Електроний ресурс] / Boychuk Vadym // Google Blogger. Режим доступу https://gorboyx88gmail.blogspot.com. Назва з екрана. Дата публікації 13.04.2023.

21. Peter F. MacNeilage, Lesley J. Rogers, Giorgio Vallortigara Evolutionary Origins of Your Right and Left Brain August 2009 Scientific American 301(1):60-7.

Boychuk V.O., Boychuk A.A., Boychuk M.V. ON THE FORMATION OF ACTIONS SEQUENCES FOR INTELLIGENT AGENTS

Modern robots and other intelligent agents are exposed to problems with the organization of autonomous existence in real environments. At the same time, the processes of organization and planning of actions, which are originally inherent to both humans and rather simple biological creatures, allow them to survive there. Therefore the paper describes the ideas of psychology and neurophysiology about the organization of formation and planning of actions sequences for humans and animals. The processes of formation and planning of actions sequences are considered on the basis of several levels of reflex rings, which allows flexible formation of actions sequences.

It is shown how this specificity can be implemented on the basis of intelligent planning systems, Transformer language model, deep artificial neural networks, spike neural networks, reinforcement learning, artificial emotion models.

Given the shortcomings of the above technologies on the basis of the bionic approach using a model based on artificial emotions, a method for forming actions sequences for intelligent agents is presented. At each moment of time, the agent executes and modifies the parameters of the new action selection policy based on the output values of the classifiers and the vector of state values.

Using the presented method, the order of chains formation of unconditional reflexes is shown. The stages of formation of conditioned reflexes on the basis of unconditioned reflexes are described. The functioning of orienting unconditional and conditioned reflexes is demonstrated for robot models in the V-REP environment.

The method is developed on the basis of analogies with the brain functioning for flexible multi-level formation of actions sequences for intelligent agents and will provide an opportunity to ensure their functioning in a complex changing environment.

Key words: *artificial intelligence, neural networks, intelligent agent, planning, artificial emotion, reinforcement learning, reflex.*