

МОДЕЛЬ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ С ПРИМЕНЕНИЕМ АДАПТИВНОЙ СЕТИ И НЕЧЕТКОГО ВЫВОДА¹

А.П. Еремеев (*eremeev@appmat.ru*)

В.С. Петров (*PetrovVS@mpei.ru*)

Национальный исследовательский университет, «МЭИ», Москва

В работе предлагается модель глубокого обучения с подкреплением (Deep Reinforcement Learning, DRL) на основе адаптивной нейронной сети и системы нечеткого вывода. Ключевой целью является разработка высокопроизводительного агента DRL, обладающего повышенным уровнем объяснимости по сравнению с традиционными агентами, функционирующими как «черные ящики». Проводится тестирование и сравнительный анализ на примере задачи CartPole (тележка с шестом). Полученные результаты подтверждают потенциал гибридных моделей, сочетающих методы DRL и нечеткой логики, для создания эффективных и интерпретируемых систем искусственного интеллекта, критически важных в прикладных областях, требующих доверия и прозрачности.

Ключевые слова: искусственный интеллект, обучение с подкреплением, нейронная сеть, нечеткая логика, нечеткие системы, система реального времени.

Введение

Традиционно в системах нечеткого управления широко применяется объединение модели динамики объекта со средствами нечеткого логического вывода, что приводит к формированию динамических нечетких регуляторов [Presup et al., 2011]. Однако в ряде прикладных задач, таких как интерактивные программные среды или виртуальная реальность, построение точной математической модели объекта затруднено или вовсе невозможно. В таких условиях классические подходы, основанные на математическом моделировании, оказываются недостаточно эффективными [Reiners et al., 2021].

¹ Работа выполнена при финансовой поддержке РФФИ (проект № 24-11-00285) <https://rscf.ru/project/24-11-00285/>.

Обучение с подкреплением (Reinforcement Learning, RL) представляет собой класс методов машинного обучения, ориентированных на выработку стратегий принятия решений в условиях неопределенности. Алгоритмы RL позволяют агенту постепенно улучшать поведение в процессе взаимодействия со средой, используя информацию о полученных наградах для корректировки будущих действий [Russell et al., 2010].

В последние годы методы глубокого обучения в сочетании с RL – методы DRL – продемонстрировали качественный прорыв: во многих задачах агенты на базе DRL показывают более высокие результаты, чем люди [Zhao et al., 2022]. Эти методы и модели на их основе успешно применяются, например: для навигации беспилотных летательных аппаратов [Walker et al., 2019], дообучения больших языковых моделей [Ouyang et al., 2022] и в других приложениях. Вместе с тем характерной проблемой таких моделей остается их ограниченная прозрачность, затрудняющая интерпретацию принятых решений.

Для устранения этого недостатка интенсивно развиваются методы и модели объяснимого искусственного интеллекта, направленные на создание моделей (и алгоритмов) с понятной и воспроизводимой логикой работы [Borys et al., 2023]. Особенно перспективными считаются гибридные нечеткие модели, поскольку они позволяют формализовать правила в виде понятных формулировок и визуализировать логические зависимости [Борисов и др., 2012], [Нечеткие гибридные системы, 2007], [Mencar et al., 2019], [Víaña et al., 2023].

Особый интерес представляют гибридные решения, объединяющие нечеткую логику и методы RL [Berenji, 1992]. Первые исследования по нечеткому RL (Q-обучению) [Er et al., 2004], [Jamshidi et al., 2015] показали эффективность в задачах управления, однако в них уделялось мало внимания вопросам прозрачности принимаемых решений. В работе [Kumar, 2020] предложен гибридный механизм, в котором базовые правила Такаги–Сугено–Канга (TSK) интегрируются с алгоритмом Deep Q-Network (DQN) [Mnih et al., 2015]. В этой схеме алгоритм RL отвечает за одновременную корректировку функций принадлежности и линейных коэффициентов вывода нечеткой модели, аналогично тому, как это делается в модели (системе) ANFIS (Adaptive Neuro-Fuzzy Inference System) [Jang, 1993], представляющей собой развитие TSK-моделей. Благодаря этому система самостоятельно подстраивает нечеткие правила под динамику окружающей среды, что подтверждается ее высокой практической эффективностью.

В предыдущих работах авторов рассматривались алгоритмы RL с нечеткой логикой и сравнивалась их эффективность [Eremeev et al., 2024], а также исследовались возможности самоорганизующейся нечеткой Q-сети [Еремеев и др., 2023], с целью применения в интеллектуальных системах поддержки принятия решений в реальном времени (ИСППР РВ).

Целью работы являются дальнейшие исследования нечетких систем с применением алгоритмов DRL в качестве решений для создания объяснимого искусственного интеллекта. Работа выполнена в рамках тематики научной группы кафедры Прикладной математики и искусственного интеллекта НИУ «МЭИ» по созданию инструментальных средств (методов, моделей, программ) конструирования ИСППР РВ для мониторинга и управления сложными техническими и другими системами.

1. Обучение с подкреплением с использованием адаптивной нейро-нечеткой модели

Интеграция в модели обучения с подкреплением и нечеткой логики была предложена [Berenji, 1992] на основе классической схемы Мамдани: при этом агент корректирует правила нечеткого вывода, получая сигналы вознаграждения за свои действия в среде. Первоначальные эксперименты, в частности, на задаче управления тележкой с шестом CartPole [Barto et al., 1983], где агент постепенно совершенствует стратегию балансирования, подтвердили практическую применимость данного подхода.

Дальнейшее развитие получило направление RL-TSK, в котором вместо простых констант в выводе правил используются аффинные функции TSK-модели. Такие гибридные модели разрабатывались для самых разных сценариев: от управления роботизированными манипуляторами до адаптации параметров в системах поддержки решений [Yan et al., 2001], а также для стратегий навигации и оптимизации ресурсов в динамических приложениях [Kumar, 2020].

В задачах управления и обучения с подкреплением одной из ключевых проблем является трудность интерпретации поведения агента. Стратегии, выработанные в процессе многократного взаимодействия со средой и оптимизации сложных критериев, обычно представляют собой черные ящики, что затрудняет их анализ и применение в критичных областях. Именно поэтому сейчас развивается направление объяснимого RL (Explainable Reinforcement Learning, XRL), фокусирующееся на создании моделей, обеспечивающих понятное объяснение принимаемых решений.

Одним из подходов в плане объяснимости в XRL является применение нечетких TSK-моделей. Благодаря возможности представлять правила в форме «если..., то...», они обладают хорошей наглядностью, а система типа универсального аппроксиматора позволяет приближать любые функции. В контексте RL это означает, что функция полезности агента $Q(s, a)$ может аппроксимироваться TSK-моделью, где каждое правило задает линейную или аффинную функцию входных переменных, а итоговой оценкой служит взвешенное по степеням истинности объединение результатов всех правил:

(1)

где s – состояние среды, a – действие агента.

Гибридная модель продемонстрировала свою работоспособность как в дискретных средах, например, в клеточных автоматах, так и в непрерывных задачах, таких как балансировка маятника, управление посадкой лунного модуля (Lunar Lander) [Barto et al., 1983].

Обучение в этой модели строится по тому же принципу, что и в классический Q-learning: текущая оценка полезности корректируется с учетом разности во времени (temporal difference, TD). Формально обновление параметров происходит по правилу

где s – текущее состояние среды, a – действие, выбранное в состоянии s ,
– следующее состояние среды после выполнения действия a ,
– вознаграждение, полученное за переход в состояние , $Q(s,a)$ – текущее значение Q-функции для пары (s,a) ,
– оценка Q-функции для следующего состояния и действия, α – скорость обучения (learning rate), $0 < \alpha \leq 1$, γ – коэффициент дисконтирования будущих вознаграждений (discount factor), $0 \leq \gamma \leq 1$

Параметры нечеткой модели можно оптимизировать с помощью модифицированной версии градиентного спуска:

$$\text{---} \quad (2)$$

где – параметры TSK-модели.

Подставляя выражение (1), получаем:

$$\text{---} \quad (3)$$

Аналогично классическим алгоритмам RL, устойчивость и эффективность обучения в нечетких алгоритмах RL могут быть значительно повышены за счет применения таких приемов, как повторное использование накопленного опыта (experience replay) и разделение функций выбора действий и оценки ценности в рамках моделей актор–критик.

Теоретическая обоснованность данного подхода базируется на известных результатах о сходимости Q-обучения при выполнении стандартных условий марковского процесса принятия решений (MDP) [Jaakkola et al., 1993]. Более того, как нейросети с различными функциями активации, так и нечеткие модели типа TSK доказали свою способность к универсальной аппроксимации. Благодаря этому TSK-модели могут эффективно заменить традиционные нейросетевые блоки в структуре DQN, сохраняя при этом фундаментальные свойства сходимости и точности аппроксимации.

Следовательно, нечеткая TSK-модель способна выполнять роль аппроксиматора функции полезности $Q(s,a)$, что подтверждает применимость алгоритма Q-обучения в сочетании с данной моделью. Это создает прочную теоретическую основу для дальнейших исследований в направлении адаптивных нечетких моделей в контексте обучения с подкреплением.

Модель (система) ANFIS, как отмечено ранее, является развитием концепции TSK-моделей и позволяет не просто задавать параметры правил вручную, а настраивать их автоматически с использованием алгоритма градиентного спуска. В отличие от подходов с жестко заданными параметрами условий (антецедентов) и следствий (консеквентов), ANFIS обеспечивает гибкую и адаптивную настройку, а также поддерживает XRL, что делает ее особенно подходящей для динамичных обучающих сред.

Для реализации заданной цели используется система ANFIS, интегрированная с нейросетевым компонентом. На первом этапе нейронная сеть извлекает информативные признаки из входных данных, после чего они преобразуются в степени принадлежности с помощью фаззификационного слоя. Далее выходы этого слоя комбинируются с условиями (антецедентами) нечетких правил путем их попарного перемножения. Полученные значения масштабируются с помощью обучаемых параметров, после чего агрегируются для формирования заключений (консеквентов) системы.

Когда требуется предсказывать несколько выходных значений одновременно (multi-output), применяется модифицированный вариант ANFIS – MANFIS (Multioutput ANFIS). В этой версии для каждого компонента выходного вектора формируется собственный набор нечетких правил, в то время как механизм дефаззификации остается идентичным классической реализации ANFIS. Схема системы ANFIS [Jang, 1993] представлена на рис. 1.

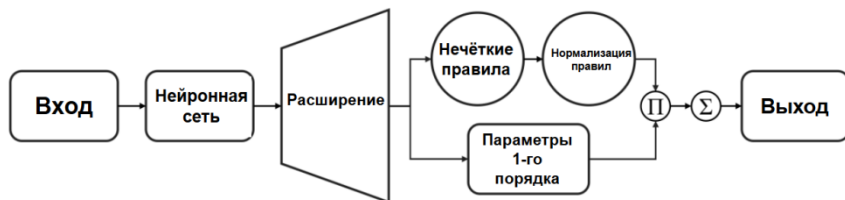


Рис. 1. Схема ANFIS

В данной модели (системе) используется нейронная сеть обобщенной структуры, параметры и глубина которой подбираются в зависимости от конкретной задачи. После первичной обработки входных данных нейронной сетью, результирующий выход трансформируется таким образом, чтобы его размер соответствовал числу заданных нечетких правил. Затем они подаются на модуль нечеткого вывода, где для каждого правила вычисляется степень его активации (firing strength), после чего значения нормализуются отдельно по каждому выходному каналу.

В случае применения ANFIS первого порядка, к каждой переменной применяется линейная функция – входное значение масштабируется с помощью обучаемого коэффициента и корректируется сдвигом (bias). Для ANFIS нулевого порядка используется упрощенная схема, в которой входной сигнал напрямую передается на следующий этап обработки, без преобразований. На заключительной стадии нормализованные коэффициенты срабатывания умножаются на соответствующие выходы от системы правил. Далее все полученные значения агрегируются путем суммирования, обеспечивая формирование итогового выхода модели с требуемыми размерностями.

2. Тестирование модели и сравнение результатов

В данном разделе приводится описание тестовой задачи CartPole, результаты моделирования и сравнение с моделями, разработанными в предыдущих работах.

2.1. Описание тестовой задачи CartPole

В задаче CartPole, называемой также задачей перевернутого маятника (inverted pendulum) (см. рис. 2) агенту необходимо управлять тележкой с прикрепленным к ней вертикальным шестом (маятником). Входной вектор среды состоит из четырех параметров: линейной скорости тележки; позиции тележки; угла отклонения шеста от вертикали; угловой скорости шеста. Действие, которое должен выбрать агент на каждом шаге, сводится к одному из двух вариантов – переместить тележку влево либо вправо.

Целью обучения является максимизация кумулятивной функции полезности для каждого состояния, что в идеале позволяет агенту накапливать наибольшее возможное вознаграждение – до 500 очков. На каждом временном шаге агент получает вознаграждение, равное 1, если система (тележка и шест) остается в допустимых пределах по всем параметрам. Таким образом, достижение общего вознаграждения в 500 свидетельствует о том, что агент успешно удерживал систему в равновесии в течение 500 последовательных шагов.

Принятие решений агентом основывается на выборе действия, соответствующего максимальному значению оценочной функции Q в текущем состоянии.

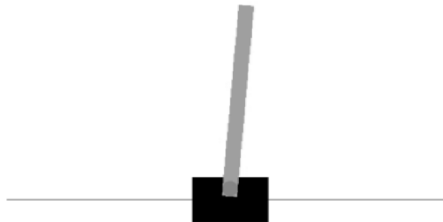


Рис. 2. Иллюстрация CartPole

2.2. Описание разработанной нечеткой модели

Нечеткие множества (довольно простые, но для решения данной задачи они подходят) для модельной задачи CartPole представлены на рис. 3.

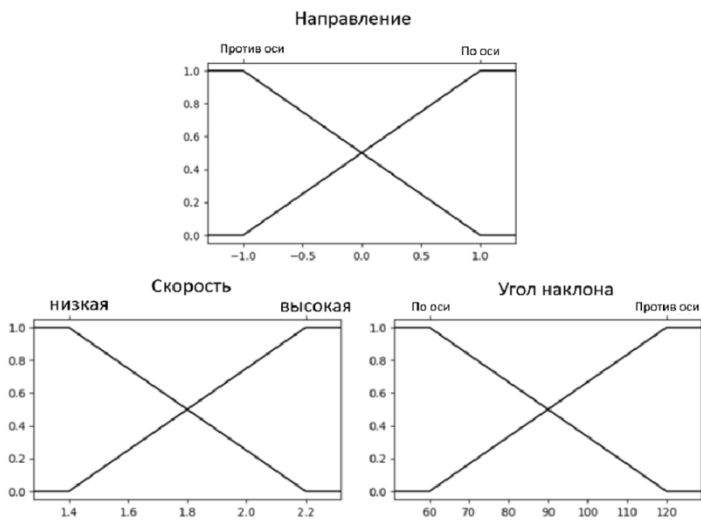


Рис. 3. Нечеткие множества для задачи CartPole

Нечеткие правила задаются следующим образом:

1. «Если направление против оси и угол наклона по оси и скорость высокая; то осевая тяга должна быть сильная, направление по оси»;
2. «Если направление против оси и угол наклона по оси и скорость низкая; то осевая тяга должна быть среднее, против оси»;
3. «Если направление тяги против оси и угол скорости против оси и скорость низкая; то осевая тяга должна быть слабая, направление против оси»;
4. «Если направление тяги против оси и угол абсолютной скорости против оси, и скорость высокая; то осевая тяга должна быть без изменений (0)»;
5. «Если направление тяги по оси и угол абсолютной скорости по оси и скорость высокая; то осевая тяга должна быть без изменений (0)»;
6. «Если направление тяги по оси и угол абсолютной скорости по оси, и скорость низкая; то осевая тяга должна быть слабая, по оси»;
7. «Если направление тяги по оси и угол абсолютной скорости против оси, и скорость низкая; то осевая тяга должна быть средняя, по оси»;

8. «Если направление тяги по оси и угол абсолютной скорости против оси, и скорость высокая; то осевая тяга должна быть сильная, по оси».

Непротиворечивость набора (базы) правил проверяется посредством выявления попарных пересечений антецедентов правил с несовпадающими консеквентами; установлено, что данный набор непротиворечив. Примеры формальной записи правил даны в работе [Еремеев и др., 2023].

Структура разработанной гибридной модели (системы) ANFIS-DQN представлена на рис. 4.

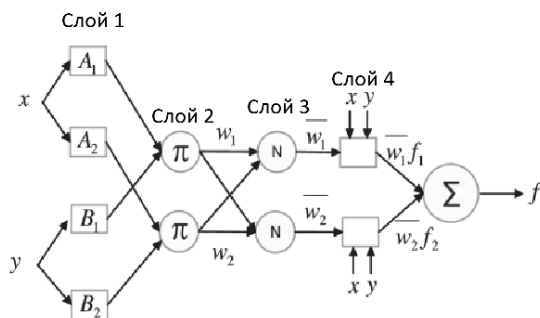


Рис. 4. Структура ANFIS-DQN

2.3. Результаты компьютерного моделирования и сравнительный анализ

Разработанная модель (система) ANFIS-DQN сравнивалась со стандартной моделью DQN, результаты приведены на рис. 5. Обе модели построены с применением механизма мягкого обновления целевой сети DQN. Количество обучаемых параметров в моделях сопоставимо: 10 290 у стандартной DQN и 10 293 в случае ANFIS-DQN. В целях обеспечения корректности сравнения обе модели обучались в идентичных условиях, с одинаковыми параметрами:

- скорость обучения (α) 0.001;
- коэффициент дисконтирования (γ) 0.99;
- размер батча для Experience Replay 128;
- число итераций 10 000.

Для оценки результатов моделей DQN и ANFIS-DQN (на рис. 5 обозначены, соответственно, как dqn и anfis) измеряется средняя награда по 10 тестовым средам каждые 2000 итераций. Цель – чтобы агент достигал средней награды 500 за 10 эпизодов.

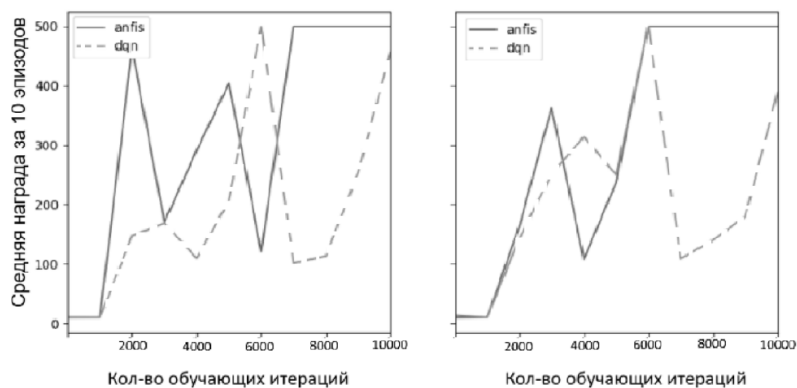


Рис. 5. Сравнение ANFIS-DQN и DQN

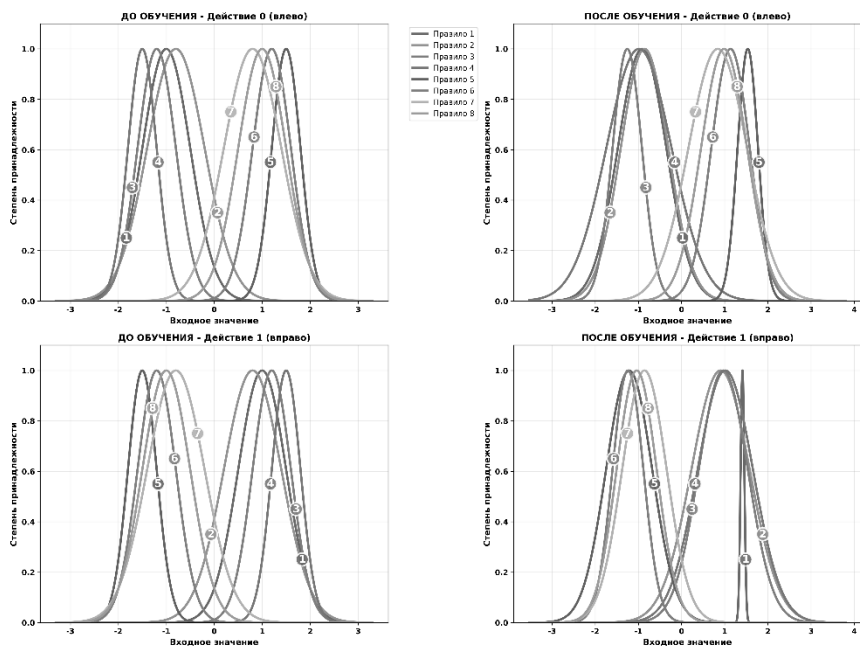


Рис. 6. Сравнение нечетки правил до и после обучения для перемещения тележки влево и вправо в среде CartPole

Из рис. 5 видно, что обе модели способны обучиться в среде, причем ANFIS-DQN обучается быстрее, чем DQN (изломы на кривых объясняются немонотонностью до определенного числа итераций соответствующих

функций оценки). Время обучения DQN – 45 секунд ANFIS-DQN – 164 секунды. По результатам моделирования можно заключить, что модель ANFIS-DQN может быть использована в ИС/ИСППР РВ. На рис. 6 приведены графики функций принадлежности нечетких правил. Видно, что некоторые кривые сместились влево или вправо. Это объясняется тем, что движение тележки в сторону, противоположную падению шеста, помогает скорректировать его положение. В работе [Еремеев и др., 2023] приведен пример тестирования на задаче «горный автомобиль» (mountain car), также показавший эффективность предложенного подхода. В дальнейшем планируется проведение экспериментов для более сложных сред (LunarLander, MuJoCo).

Заключение

В рамках данного исследования была разработана модель (система) ANFIS-DQN на основе DQN-модели обучения с подкреплением, дополненный адаптивной нейро-нечеткой системой вывода ANFIS, поддерживающая объяснимое RL. Разработанная модель (система) ANFIS-DQN прошла тестирование на задаче CartPole и продемонстрировала способность к успешному обучению. Выполнено сравнение ANFIS-DQN с классической реализацией DQN. Результаты показали, что нейро-нечеткая модель ANFIS-DQN обладает лучшей стабильностью и более быстрой сходимостью. В отличие от стандартных нейросетевых моделей, которые функционируют как «черный ящик», модель на базе нечетких правил обеспечивает интерпретируемость поведения как до, так и после обучения. Выполненная работа является частью комплексной задачи по созданию интегрированной среды для разработки перспективных ИС/ИСППР РВ.

Список литературы

- [Борисов и др., 2012] Борисов В.В., Крутлов В.В., Федулов А.С. Нечеткие модели и сети. – 2-е изд., стереотип. – М.: Горячая линия-Телеком, 2012. – 284 с. – ISBN 978-5-9912-0283-1.
- [Еремеев и др., 2023] Еремеев А.П., Сергеев М.Д., Петров В.С. Интеграция методов обучения с подкреплением и нечеткой логики для интеллектуальных систем реального времени // Программные продукты и системы. – 2023. – № 4. – С. 600-606. – DOI: 10.15827/0236-235X.144.600-606.
- [Нечеткие гибридные системы, 2007] Батыршин И.З., Недосекин А.О., Стецко А.А., Тарасов В.Б. Нечеткие гибридные системы. Теория и практика: учебное пособие / под ред. Н.Г. Ярушкиной. – М.: Физматлит, 2007. – 208 с. – ISBN 978-5-9221-0786-0.
- [Barto et al., 1983] Barto A.G., Sutton R.S., Anderson C.W. Neuronlike adaptive elements that can solve difficult learning control problems // IEEE Transactions on Systems, Man, and Cybernetics. – 1983. – Vol. 13(5). – P. 834-846. – doi: 10.1109/TSMC.1983.6313077.

- [**Berenji, 1992**] Berenji H.R. A reinforcement learning-based architecture for fuzzy logic control // *International Journal of Approximate Reasoning*. – 1992. – Vol. 6(3). – P. 267-292. – doi: 10.1016/0888-613X(92)90021-C
- [**Borys et al., 2023**] Borys K., Schmitt Y.A., Nauta M., Seifert C., Krämer N., Friedrich C.M., Nensa F. Explainable AI in medical imaging: an overview for clinical practitioners – beyond saliency-based XAI approaches // *European Journal of Radiology*. – 2023. – Vol. 162(11). – P. 780-786. – doi:10.1016/j.ejrad.2023.110786.
- [**Er et al., 2004**] Er M.J., Deng C. Online tuning of fuzzy inference systems using dynamic fuzzy Q-learning // *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*. – 2004. – Vol. 34(3). – P. 1478-1489. – doi: 10.1109/TSMCB.2004.826353.
- [**Eremeev et al., 2024**] Eremeev A.P., Sergeev M.D., Petrov V.S. Integrating reinforcement learning methods with neural networks and fuzzy logic in real-time intelligent systems // In: 2024 7th International Conference on Information Technologies in Engineering Education (Inforino), Moscow, Russian Federation, 2024. – P. 1-4. – doi: 10.1109/Inforino60363.2024.10551979.
- [**Jaakkola et al., 1993**] Jaakkola T., Jordan M., Singh S. Convergence of stochastic iterative dynamic programming algorithms // In: *Advances in Neural Information Processing Systems*. – 1993. – Vol. 6(1). – P. 1-8. – doi: 10.5555/297645.297652.
- [**Jamshidi et al., 2015**] Jamshidi P., Sharifloo A.M., Pahl C., Metzger A., Estrada G. Self-learning cloud controllers: fuzzy Q-learning for knowledge evolution // In: 2015 International Conference on Cloud and Autonomic Computing. – IEEE, 2015. – P. 208-211. – doi: 10.1109/ICCAC.2015.38.
- [**Jang, 1993**] Jang J.S.R. ANFIS: adaptive-network-based fuzzy inference system // *IEEE Transactions on Systems, Man, and Cybernetics*. – 1993. – Vol. 23(3). – P. 665-685. – doi: 10.1109/21.256541.
- [**Kumar, 2020**] Kumar S. Learning of Takagi-Sugeno fuzzy systems using temporal difference methods // Redmond: DigiPen Institute of Technology. – 2020. – 65 p.
- [**Mencar et al., 2019**] Mencar C., Alonso J.M. Paving the way to explainable artificial intelligence with fuzzy modeling. In: *Fuzzy Logic and Applications*. – Springer, 2019. – P. 215-227. – doi: 10.1007/978-3-030-19591-5_19.
- [**Mnih et al., 2015**] Mnih V., Kavukcuoglu K., Silver D., Rusu A.A., Veness J., Bellemare M.G., Graves A., Riedmiller M., Fidjeland A.K., Ostrovski G., Petersen S., Beattie C., Sadik A., Antonoglou I., King H., Kumaran D., Wierstra D., Legg S., Hassabis D. Human-level control through deep reinforcement learning // *Nature*. – 2015. – Vol. 518(7540). – P. 529-533. – doi: 10.1038/nature14236.
- [**Ouyang et al., 2022**] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P., Zhang C., Agarwal S., Slama K., Ray A., Schulman J., Hilton J., Kelton F., Miller L., Simens M., Askell A., Welinder P., Christiano P.F., Leike J., Lowe R.J. Training language models to follow instructions with human feedback // *Advances in Neural Information Processing Systems*. – 2022. – Vol. 35. – P. 27730-27744. – doi: 10.48550/arXiv.2203.02155.
- [**Precup et al., 2011**] Precup R.E., Hellendoorn H. A survey on industrial applications of fuzzy control // *Computers in Industry*. – 2011. – Vol. 62(3). – P. 213-226. – doi: 10.1016/j.compind.2010.10.001.

- [Reiners et al., 2021]** Reiners D., Davahli M.R., Karwowski W., Cruz-Neira C. The combination of artificial intelligence and extended reality: a systematic review // *Frontiers in Virtual Reality*. – 2021. – Vol. 2(1). – P. 1-24. – doi: 10.3389/frvir.2021.72193.
- [Russell et al., 2010]** Russell S.J., Norvig P. *Artificial Intelligence: A Modern Approach*. – M.: Pearson Education, Inc., 2010.
- [Viana et al., 2023]** Viana J., Ralescu S., Cohen K., Ralescu A., Kreinovich V. Extension to multidimensional problems of a fuzzy-based explainable and noise-resilient algorithm // In: *Decision Making Under Uncertainty and Constraints: A Why-Book*. – Springer, 2023. – P. 289-296. – doi: 10.1007/978-3-031-27402-2_26.
- [Walker et al., 2019]** Walker O., Vanegas F., Gonzalez F., Koenig S. A deep reinforcement learning framework for UAV navigation in indoor environments // In: *2019 IEEE Aerospace Conference*. – IEEE, 2019. – P. 1-14. – doi: 10.1109/AERO.2019.8742226.
- [Yan et al., 2001]** Yan X., Deng Z., Sun Z. Competitive Takagi-Sugeno fuzzy reinforcement learning // In: *Proceedings of the 2001 IEEE International Conference on Control Applications (CCA'01) (Cat. No. 01CH37204)*. – IEEE, 2001. – P. 878-883. doi: 10.1109/CCA.2001.948137
- [Zhao et al., 2022]** Zhao E., Yan R., Li J., Li K., Xing J. AlphaHoldem: High-performance artificial intelligence for heads-up no-limit poker via end-to-end reinforcement learning // In: *Proceedings of the AAAI Conference on Artificial Intelligence*. – AAAI, 2022. – Vol. 36(4). – P. 210-218. – doi: 10.1609/aaai.v36i4.20394.