Entanglement Learning for Adaptive AI

Solution Technical Details

Table of Contents

1 What is Entanglement Learning?

Adaptability in AI systems hinges on their ability to identify performance deviations, which requires comparing actual and expected performance values [6]. Current AI systems rely on designers to provide the expected or reference value, necessitating human intervention when decisions, predictions, or tasks deviate from expectations [27, 25]. This dependency limits AI systems' capacity for self-reflection and autonomous gap quantification [26]. Entanglement Learning (EL) addresses this limitation by providing AI systems with an intrinsic reference, their entanglement with their environment [12], enabling them to quantify performance gaps due to changes and generate control signals to adjust their performance without external intervention. This self-reflective capability enhances AI systems' adaptability and autonomy in dynamic environments.

2 The Technical Solution

Entanglement Learning (EL), a framework designed to enhance the learning, adaptability, and autonomy of Reinforcement Learning (RL) agents, with an initial focus on model-based RL. A model-based RL agent learns to predict state transitions and rewards from state-action-next state sequences, creating a dynamic model of the environment based on its actions [32]. The agent uses this internal model to plan and continually optimize its policy, which maps states to actions to maximize cumulative rewards over time and effectively achieve its objectives [21]. The core research hypothesis of EL is that entanglement, an information-theoretic metric that captures the degree of mutual predictability between events [12], directly correlates to the degree of alignment between an agent's objectives and its environment.

That is, the greater the entanglement within the state-action-next state process, the higher the predictability and control the agent has over its actions and environment. This increased control indicates the agent's potential and ability to achieve its objectives and maintain adaptability and resilience in dynamic environments. Accordingly, maintaining and increasing higher levels of entanglement would enable higher performance, adaptability, and resilience.

To achieve this adaptability and resilience, the proposed technical solution introduces an additional learning layer to existing RL agents, the “EL Framework” (Figure 1). The EL framework consists of two main components: the Sematic Matrix (SMX) and the Entanglement Controller (EC). **The Semantic Matrix (SMX)** is primed by capturing the conditional probabilities of state-action-next state as learned from historical records and later as observed during the agent-environment interaction. The SMX provides various entanglement metrics related to state-action pairs distributions and changes over time. The second main component is the **Entanglement Controller (EC)** then relies on the provided entanglement metrics (in bits) and the agent objectives to generate adaptive control signals in real time. The control signals are integrated with the RL algorithm to modulate its hyperparameters, reward function, Bellman equation, and ultimately its policy.

The Sematic Matrix (SMX) main product is the **Information Genome (InfoGen)**, which is a structured, task- and environment-specific representation of the learned information patterns and predictability between an agent's actions and the states of its environment for a specific task. It encapsulates the domain-specific information measures—or entanglements—that the agent has learned through its interactions towards a specific task. Serving as a guiding blueprint, the InfoGen provides the agent with baseline estimates of probable actions in response to observed states and anticipated state changes following specific actions.

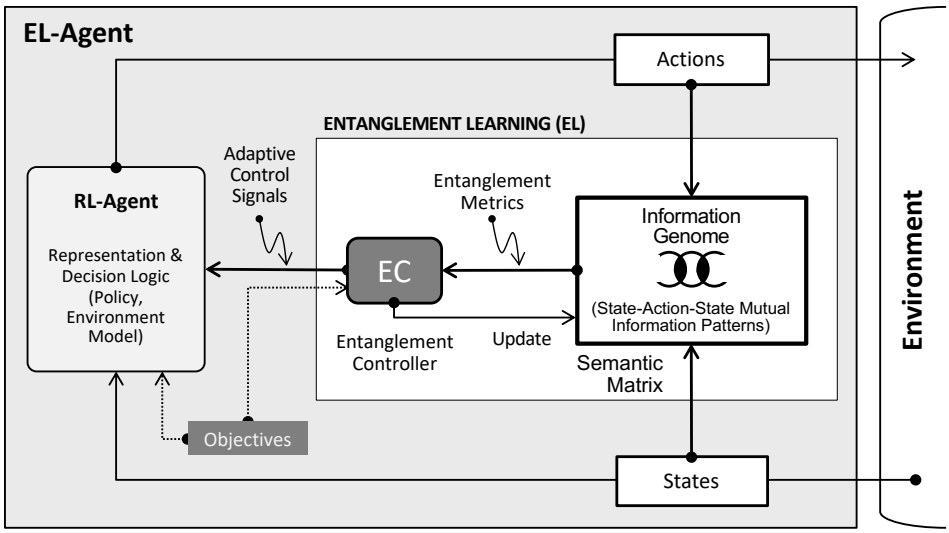

Figure 1. Entanglement Learning Architecture Overview

The agent can then rely on the InfoGen to simulate possible state-action-next state entanglement values and relies on the outcomes to adjust its actions towards higher entanglement. In addition, the InfoGen is transferable across agents, or multiple such genomes can be integrated to expand an agent's task scope within a specific domain.

3 Entanglement Learning Main Algorithm

The following pseudo EL algorithm not only outlines the operational steps of an EL-based agent, but also reflects the research approach to develop and validate the EL Framework:

	Steps
1	**Initialization**: Load the Information Genome (InfoGen) with historical data or data from the end of the learning phase.
2	**State Observation**: The agent observes the current state from the environment.
3	**Action Selection**: The agent's policy suggests an action based on the observed state.
4	**Expected EL Calculation**: Retrieve expected next states and their entanglement values from InfoGen for the recommended action.
5	**Action Execution**: The agent executes the recommended action.
6	**Action Update**: Adjust action selection towards higher entanglement by incorporating feedback from InfoGen.
7	**State Transition and Entanglement Evaluation**: Observe the actual next state and calculate the change in entanglement and asymmetry from InfoGen.
8	**Next State Analysis**: InfoGen provides possible subsequent actions and their entanglement values for the observed next state.
9	**Entanglement Optimization**: Choose the next action to compensate for the entanglement change, either increasing overall entanglement or restoring asymmetry.
10	**RL Agent Update**: Update the policy and/or environment model based on entanglement changes to improve decision-making.

11	**InfoGen Update**: Update the InfoGen entropies based on actual states and actions.
12	**Repeat**: The cycle repeats with the new state and selected action, continually adjusting based on new data and entanglement metrics.

Table 1. Entanglement Learning Pseudo Algorithm.

By providing a principled, agent-intrinsic, and data-driven approach to adaptability, EL has the potential to revolutionize the field of RL and autonomous systems in general. The proposed technical solution is grounded in information theory [29,3], RL methodologies [22], meta-learning concepts [15], and adaptive algorithms [20]. It also leverages the latest advancements in machine learning, particularly in the areas of automated learning process control [13], and dynamic system adaptation [1].

4 State of the Art and Challenges in Reinforcement Learning (RL)

Reinforcement Learning (RL) has emerged as a powerful framework for enabling agents to learn and make decisions in complex environments. RL agents learn through interaction with their environment, receiving rewards or penalties for their actions, and updating their policies to maximize cumulative rewards over time [32]. This approach has achieved remarkable successes in various domains, including game playing [31], robotics [19], and autonomous systems [18].

The challenges faced by reinforcement learning (RL) algorithms can be largely attributed to the uncertainties and complexities in learning the state-action-next state relationships. At the core of RL is the goal of learning optimal policies that map states to actions, maximizing cumulative rewards [4]. However, managing the uncertainties associated with the state-action-next state transitions pose significant challenges. First, the stochasticity and partial observability of the environment introduce uncertainties in predicting the next state given a current state and action [5]. Second, the curse of dimensionality, where the state and action spaces grow exponentially, makes it challenging to learn and represent the full state-action-next state relationships [2]. Furthermore, the non-stationarity of the environment, where the underlying dynamics may change over time, adds another layer of uncertainty in estimating the long-term consequences of actions [23]. These uncertainties, and the very nature of RL in learning through interactions, make it challenging for RL agents to learn accurate models of the environment with limited data sets, generalize to unseen states, and adapt to changing conditions. Addressing these uncertainties and learning reliable state-action-next state relationships is crucial for developing robust and efficient RL algorithms that can tackle real-world problems [17].

5 Entanglement Learning (EL) Overview

Entanglement Learning (EL) introduces a novel perspective on addressing the challenges and uncertainties in reinforcement learning (RL) by drawing inspiration from communication theory. EL views the state-action-next state process performed by an agent as it interacts with its environment as a communication process between a source (set of agent's available actions) and a destination (set of environment's observed states) [9,11], (Figure 2).

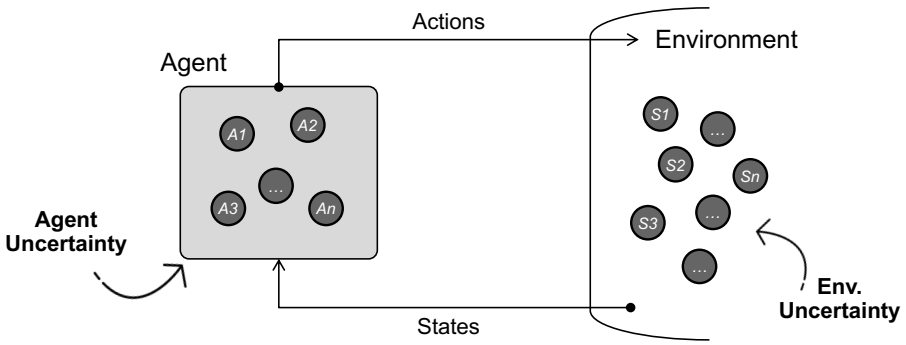

Figure 2. Agent-Environment interaction as a process of communication between the agent defined by its actions, and the environment, defined by the various task states.

However, and unlike communication theory, which assumes symmetric exchanges between a source and destination using the same signals along a shared channel [29,3]. EL posits that communication in agent-environment interactions is inherently asymmetric, as the state-action-next state exchanges occur through two distinct channels: the state-action within the agent "channel" and the action-next state through the environment "channel", which results in two sources of uncertainties instead of one. By considering the two uncertainties in the state-action-next state relationships, EL leverages information-theoretic concepts to quantify and manage these uncertainties [10,12].

Central—and novel—to EL is capturing the two uncertainties, which are tightly corelated, as one concept: entanglement [12]. **Agent-environment entanglement is thus provided by calculating the agent-environment mutual information over the complete interaction cycle. Accordingly, we define entanglement as the mutual information between the observed state-selected agent action and the resulting next observed state**. Agent-environment entanglement, or information, thus provides the level of mutual predictability along the states-actions-next states process.

6 Entanglement Learning Research Hypothesis

In traditional reinforcement learning (RL), the reward function serves as the independent variable, defining the criteria for successful behavior by providing immediate feedback to the agent based on its actions. The reward function is set by the task designer and influences the behavior and learning trajectory of the agent [30]. The dependent variable in RL is the agent's performance with respect to the use case objectives, which can be measured in various ways, such as efficiency, accuracy, or success rate in achieving the defined goals [14]. The agent's performance is directly influenced by how it responds to the reward function [16].

The core hypothesis of Entanglement Learning (EL) asserts that enhancing the bidirectional predictability—quantified as mutual information or entanglement—between an agent and its environment boosts performance metrics such as learning speed, adaptability, reliability, and transparency. By measuring and actively managing this entanglement, EL aims to demonstrate superior, sustainable performance and knowledge transfer compared to traditional methods focused on static rewards and exploratory behaviors.

7 Quantifying Agent-Environment Interaction Uncertainties using Information Theory

In order to calculate the two components of the entanglement: **MIa(A;S)** the state-action mutual information (the lower case "a" indicates that this information is related to the agent as a channel,) and **MIv(S';A)** the action-next state mutual information (the lower case "v" indicates that this information is related to the environment channel,) we consider the set of all actions available to the agent (A) and the set of all observable states by the agent (S) (Figure 3).

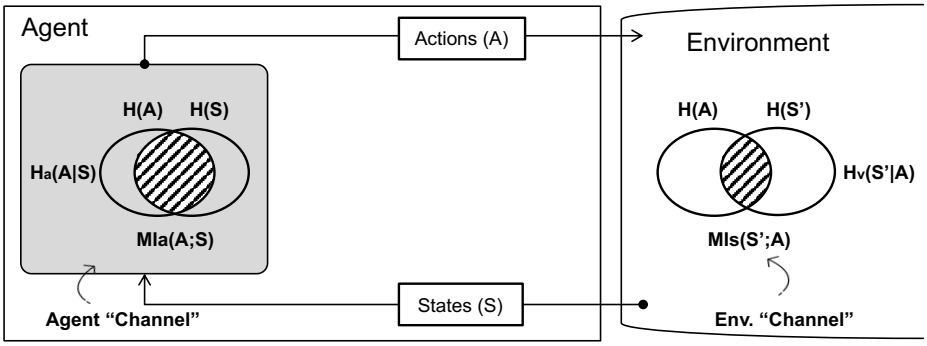

Figure 3. Agent-Environment interaction uncertainties and mutual information. The use of "S'" is merely to indicate the next state.

According to communication theory, we then have the following quantities to calculate the involved mutual information in along the agent-environment interactions:

- H(A) is the actions' entropy, the uncertainty about the agent actions. The number of bits the agent requires to select an action out of the set of actions. That is, if the actions' entropy = 4 bits, and assuming equally distributed actions, then the agent has 16 actions to choose from.
- H(S) is the states' entropy; the agent's uncertainty about the observed states, the number of bits it requires to identify a state from the set of observable states.
- Ha(A|S) is the uncertainty about selecting an action "A", given an observed state "S". It is the number of bits required to select an action knowing a state. The lower the value, the higher the predictability of the action, the higher, the less predictable is the action. The "a" indicates this uncertainty is associated with the agent "channel".
- Hv(S'|A) is the uncertainty about the resulting next state "S", given action "A". It is the number of bits required to predict a state after the agent takes an action A. The lower, the higher the predictability of the next state, the higher, the less the predictability of which state might result when taking action A. The "v" indicates that this value is associated with the environment "channel".
- MIa (A;S), is the state-actions mutual information. It captures the interdependence between the environment's states and the agent's chosen actions. It is the number of bits shared between the actions and the states, the higher, the higher the predictability of an action given a state. According to communication theory, the agent "channel" mutual information is given by:

$$MIa(A;S) = H(A) - H(A|S)$$

- MIv (S';A), is the actions-states mutual information. It captures the interdependence between the agent's actions and the resulting state of the environment. It is the number of bits shared between the states and

the actions, the higher, the higher the predictability of a state given an action. In a deterministic environment, this value is at maximum. The mutual environment channel mutual information is calculated based on the environment channel related entropies:

$$MIv(S';A) = H(S') - H(S'|A)$$

8 Entanglement Metrics

The entanglement metrics introduced in EL, such as entanglement strength (ψ), differential entanglement (Δψ), and entanglement asymmetry (Λψ), are fundamentally new and grounded in a novel definition of information as a measure of entanglement. This innovative perspective on information theory sets EL apart from traditional approaches and enables a deeper understanding of the complex dynamics between agents and their environments [8].

8.1 Entanglement (ψ, psi):

We define entanglement here as the mutual predictability between the agent and the environment along the closed-loop communication cycle between the two, or in Reinforcement learning notation, along the state-action-next state process. Given the states-actions and actions-states mutual information, we define entanglement as:

$$\psi = MI\ (S,A;S')$$

That is, entanglement, in bits, equals the mutual information between a state-action pair and the resulting next state. It indicates the predictability, and interdependence, between the agent and the environment. It quantifies the degree of overall mutual dependency between the agent and its environment.

8.2 Differential Entanglement (Δψ, delta psi)

$$\Delta\psi = \psi(t) - \psi(t\text{-}1)$$

- ψ(t) represents the entanglement metric value at the current time step t.
- ψ(t-1) represents the entanglement metric value at the previous time step t-1.

The differential entanglement Δψ quantifies the change in entanglement between consecutive time steps. It measures how the mutual predictability and coupling between the agent's actions and the environment's states evolve over time. According to this definition, the value of Δψ can provide the following insights about the agent-environment interactions:

Positive Δψ (Δψ > 0): A positive value of Δψ indicates an increase in entanglement from the previous time step to the current time step. This suggests that the mutual predictability and coupling between the agent's actions and the environment's states have strengthened. It implies that the agent's actions have become more informative about the environment's states, or the environment's states have become more predictable based on the agent's actions, compared to the previous time step. A consistently positive Δψ over time indicates a trend of increasing entanglement, suggesting that the agent is learning and adapting to the environment effectively.

Negative Δψ (Δψ < 0): A negative value of Δψ indicates a decrease in entanglement from the previous time step to the current time step. This suggests that the mutual predictability and coupling between the agent's actions and the environment's states have weakened. It implies that the agent's actions have become less informative about the environment's states, or the environment's states have become less predictable through the agent's actions, compared to the previous time step. A consistently negative Δψ over time indicates a trend of decreasing entanglement, suggesting that the agent may be struggling to

learn, adapt to, or predict its environment effectively.

Zero Δψ (Δψ = 0): A zero value of Δψ indicates no change in entanglement from the previous time step to the current time step. This suggests that the mutual predictability between the agent's actions and the environment's states have remained stable. It implies that the agent's learning and adaptation to the environment have reached a steady state, or the environment's dynamics have not significantly changed.

Δψ Magnitude: The magnitude of Δψ reflects the extent of the change in entanglement. Larger values of Δψ indicate more significant changes in the mutual predictability and coupling between the agent's actions and the environment's states. By monitoring the values of Δψ over time, we can gain insights into the dynamics of the agent's learning and adaptation process. Consistently positive Δψ values suggest effective learning and increasing entanglement, while consistently negative Δψ values may indicate challenges in learning or a need for adjustments in the agent's strategy. Fluctuations in Δψ can reveal patterns of exploration and exploitation, as well as the agent's responsiveness to changes in the environment. The differential entanglement Δψ provides a valuable metric for assessing the progress and quality of the agent's learning and adaptation in the context of Entanglement Learning. It offers a dynamic perspective on how the entanglement between the agent and the environment evolves over time, regardless of its rewards, thus enabling insights into the effectiveness of the learning process and potential areas for improvement.

8.3 Entanglement Asymmetry (Λψ, lambda psi)

$$\Lambda\psi = MIa(A;S) - MIv(S';A)$$

Entanglement asymmetry Λψ measures the imbalance or directionality in the mutual predictability and coupling between the agent's actions and the environment's states. It quantifies the extent to which one direction of influence dominates the other. Similar to the differential entanglement, the value of Λψ can provide different insights about the agent-environment interactions uncertainty and dynamics:

Positive Λψ (Λψ > 0): A positive value of Λψ indicates that the mutual information of the agent's actions given the environment's states (MIa) is greater than the mutual information of the environment's states given the agent's actions (MIv). This suggests that the agent's actions are more predictable and influenced by the environment's states than vice versa. It implies that the agent has a stronger understanding of how the environment's states affect its actions, i.e., which action is to select for which state, allowing for more informed decision-making and indicating an effective policy. A consistently positive Λψ over time suggests that the agent is effectively utilizing the information from the environment's states to guide its actions.

Negative Λψ (Λψ < 0): A negative value of Λψ indicates that the mutual information of the environment's states given the agent's actions (MIv) is greater than the mutual information of the agent's actions given the environment's states (Mia). This suggests that the environment's states are more predictable and influenced by the agent's actions than vice versa. It implies that the agent's actions have a stronger impact on shaping the environment's states, potentially indicating a higher level of control or influence over the environment, and an effective environment model. A consistently negative Λψ over time suggests that the agent is effectively shaping the environment's states through its actions.

Zero Λψ (Λψ = 0): A zero value of ΔMI indicates perfect symmetry in the mutual predictability and coupling between the agent's actions and the environment's states. This suggests that the agent's actions and the environment's states are equally informative about each other, and there is no

dominant direction of influence. It implies a balanced and reciprocal relationship between the agent and the environment, where both entities have an equal impact on each other.

Λψ Magnitude: The magnitude of Λψ reflects the extent of the asymmetry or imbalance in the entanglement between the agent and the environment. Larger absolute values of Λψ indicate a more significant dominance of one direction of influence over the other.

Changes of Λψ values over time provide insights into the evolving dynamics, stability, and resilience of the agent-environment interaction. Consistently positive Λψ suggest that the agent is effectively leveraging the states to inform its actions, while consistently negative Λψ indicate that the agent's actions are significantly shaping the environment's states. Fluctuations in Λψ can reveal shifts in the balance of influence between the agent and its environment, potentially indicating changes in the agent's strategies or the environment's dynamics. A consistently growing Λψ can indicate a developing imbalance in the agent-environment predictability.

8.4 Using the Entanglement Metrics to Control Agent Behavior

The entanglement metrics and their evaluations are the key metrics that the agent has to capture and analyze for implementing steps 8 and 9 of the EL algorithm, as outlined in Table 1.

For example, the entanglement metric, ψ, can be used to shape the agent's rewards, steering it towards actions that enhance entanglement with the environment. Incorporating ψ into the reward function focuses the agent on highly informative state-action pairs, accelerating learning by minimizing unnecessary exploration [7]. Differential entanglement, Δψ, can be used adjust the agent's policy in response to changing entanglement dynamics; a positive Δψ encourages exploiting current strategies, while a negative Δψ prompts exploration of new tactics. This dynamic policy adaptation allows the agent to respond to evolving entanglement patterns and maintain a balance between exploration and exploitation [33]. Entanglement asymmetry, Λψ, can be used to update the agent's model to better understand environmental dynamics or state-action values, depending on whether the agent's actions or the environment's states dominate, thus ultimately improve its understanding of the environment [21].

Additionally, integrating entanglement metrics into the Bellman equation enhances RL algorithms by aligning value estimates and action selection with both expected rewards and the information-theoretic properties of agent-environment interactions. This methodological adaptation ensures that decisions not only aim for maximum rewards but also address the inherent uncertainties in the environment, fostering more effective learning and adaptation strategies. This comprehensive approach leverages ψ, Δψ, and Λψ to refine the agent's operational framework, optimize performance, and maintain adaptability through informed exploratory and exploitative actions.

9 Providing Entanglement Values Based on Historical Records

To apply Entanglement Learning (EL) and define the task-specific Information Genome, InfoGen, we will identify and collect data from healthcare and finance domains, such as patient records [24] and financial time series [28]. The data will be preprocessed and binned to define discrete states and actions [5]. For healthcare, for example, states are represented by patient conditions, and actions represent treatments. In finance, states are market conditions, and actions are trading decisions. The actions and states data are represented as discrete features, i.e. binned into specific predefined discrete ranges, and captured into the Semantic Matrix (SMX) (Figure 1).

The SMX is a central computational component of the Entanglement Learning framework that, based on

the distribution of the various features, calculates the conditional probabilities of state-action pairs observed during the agent-environment interaction and their associated entropies. The result is the task-specific Information Genome, or InfoGen. The SMX updates the InfoGen with each interaction and provides the various entanglement metrics: ψ, Δψ, Λψ (step 4 and 7 of the EL algorithm, Table 1). The InfoGen serves as a compact representation of the agent entire interaction history and accordingly enable the calculation of entanglement metrics and the discovery of patterns and dependencies in the agent's behavior to adjust the agent's policy /environment model for effective learning and adaptation.

10 Bibliography

1. Åström, Karl Johan, and Boris Wittenmark. "Adaptive control." Courier Corporation, 2013.
2. Bellman, R. (1957). Dynamic Programming. Princeton University Press.
3. Cover, T. M., & Thomas, J. A. (2012). Elements of information theory. John Wiley & Sons.
4. Ding, Z., Huang, Y., Yuan, H., & Dong, H. (2020). Introduction to reinforcement learning. Deep reinforcement learning: fundamentals, research and applications, 47-123.
5. Dulac-Arnold, G., et al. (2019). Challenges of real-world reinforcement learning. Proceedings of the 36th International Conference on Machine Learning, 97, 1711-1720. (arXiv:1904.12901v1).
6. Ghandar, A., & Michalewicz, Z. (2020). Adapting to changing environments: A survey of adaptive systems and control. IEEE Access, 8, 166077-166095.
7. Guo, Z. D., et al. (2019). Efficient exploration with self-imitation learning via trajectory-conditioned policy. arXiv preprint arXiv:1907.10247.
8. Hafez W. (2019) "Human Digital Twin—Enabling Human-Multi Smart Machines Collaboration." Intelligent Systems and Applications, IntelliSys 2019. DOI:10.1007/978-3-030-29513-4_72.
9. Hafez, W. (2010) "Intelligent System–Environment Interaction as a Process of Communication." Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 2010. DOI:10.1109/ICSMC.2010.5642335.
10. Hafez, W. (2012) "Approach for Defining Intelligent Systems Technical Performance Metrics." Proceedings of the Workshop on Performance Metrics for Intelligent Systems, PerMIS'12. DOI:10.1145/2393091.2393100.
11. Hafez, W. (2012) "Development of a Framework for Measuring Cognitive Process Performance." Biologically Inspired Cognitive Architectures (BICA) 2012. DOI:10.1007/978-3-642-34274-5_29.
12. Hafez, W. (2023) "Information as Entanglement—A Framework for Artificial General Intelligence." In: Goertzel, B., et. al. (eds) Springer LNCS, AGI 2022. Doi: 10.1007/978-3-031-19907-3_3.
13. He, P. and Jagannathan, S., "Reinforcement learning-based output feedback control of nonlinear systems with input constraints," Proceedings of the 2004 American Control Conference, Boston, MA, USA, 2004, pp. 2563-2568 vol.3, doi: 10.23919/ACC.2004.1383851.
14. Henderson, P., et al. (2018). "Deep reinforcement learning that matters." Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
15. Hospedales, Timothy M., et al. "Meta-learning in neural networks: A survey." arXiv preprint arXiv:2004.05439 (2020). https://arxiv.org/abs/2004.05439.
16. Jaderberg, M., et al. (2017). Reinforcement learning with unsupervised auxiliary tasks. arXiv preprint arXiv:1611.05397.
17. Khetarpal, K., et al. (2020). Towards continual reinforcement learning: A review and perspectives. arXiv preprint arXiv:2012.13490.
18. Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Sallab, A. A. A., Yogamani, S., & Pérez, P. (2021). Deep
19. Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. The International Journal of Robotics Research, 32(11), 1238-1274.
20. Komorowski, Marcin, et al. "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care." Nature medicine 24.11 (2018): 1716-1720. https://www.nature.com/articles/s41591-018-0213-5.
21. Moerland, T. M., et al. (2020). Model-based reinforcement learning: A survey. arXiv preprint arXiv:2006.16712.

22. Nguyen, T., Nguyen, N., and Nahavandi, S., "Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications," in *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826-3839, Sept. 2020, doi: 10.1109/TCYB.2020.2977374.
23. Padakandla, S. (2020). A survey of reinforcement learning algorithms for dynamically varying environments. ACM Computing Surveys (CSUR), 53(5), 1-38.
24. Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 18.
25. Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature.
26. Schmill, M. D., Oates, T., & Cohen, P. R. (1999). Learning from self-reflection: A system architecture. AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems.
27. Seshia, S. A., & Sadigh, D. (2016). Towards verified artificial intelligence. arXiv preprint arXiv:1606.08514.
28. Sezer, O. B., et al. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing, 90, 106181.
29. Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of Communication. Urbana, IL: The University of Illinois Press, 1-117.
30. Silver, D., et al. (2021). Reward is enough. Artificial Intelligence, Volume 299, 2021, 103535, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2021.103535.
31. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.
32. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT Press.
33. Zhaohan Daniel Guo and Emma Brunskill, Directed Exploration for Reinforcement Learning. 2019, arXiv:1906.07805.