1989: Q-Learning (QL):

- Ziel: Schätzung der optimalen Aktions-Wert-Funktion (Q-Funktion)
- Vorgehen: Q-Werte werden auf Grundlage der Bellman-Gleichung aktualisiert. Dabei kommt eine Greedy-Policy zur Aktionenauswahl zum Einsatz
- Einsatzbereiche: Fragestellungen mit diskretem Zustands- und Aktionsraum
- Beispielhafte Gym-Umgebung: „FrozenLake-v0 “

1992: REINFORCE

- Ziel: Optimierung der Policy durch Gradientensteigung
- Vorgehen: Nutzung der Rückkehr aus Episoden, um die Policy zu verbessern
- Einsatzbereiche: Fragestellungen mit diskreten oder kontinuierlichem Aktionsraum
- Beispielhafte Gym-Umgebung: „CartPole-v1“

2000: Policy Gradient (PG):

- Ziel: Optimierung der Policy durch Gradientensteigung
- Vorgehen: Schätzen des Gradienten der erwarteten Belohnung in Bezug auf die Policy Paramter. Aktualisieren der Policy in Richtung des steigenden Gradienten
- Einsatzbereiche: Fragestellungen mit diskreten oder kontinuierlichem Aktionsraum
- Beispielhafte Gym-Umgebung: „MountainCarContinous-v0“

2005: Neural Fitted Q-Iteration (NFQ):

- Ziel: Schätzung der Q-Funktion durch ein neuronales Netz
- Vorgehen: Nutzung von Batch-Updates und einem neuronalen Netzwerk zur Funktionsschätzung
- Einsatzbereiche: Fragestellungen mit kontinuierlichem Zustands- und diskretem Aktionsraum
- Beispielhafte Gym-Umgebung: „MountainCarContinous-v0“

2013: Deep Q-Network (DQN):

- Ziel: Verknüpfung von Q-Learning mit tiefen neuronalen Netzen
- Vorgehen: Einsatz eines tiefen neuronalen Netzwerks zur Schätzung der Q-Funktion und Stabilisierung des Lernvorgangs durch die Wiederholung von Erfahrungen
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem Aktionsraum
- Beispielhafte Gym-Umgebung: „Highway-v0“

2015: Dueling Double Deep Q-Network (DDQN):

- Ziel: Stabilitäts- und Effizienzsteigerung von DQNs durch getrennte Schätzung des Zustandswerts und der Vorteile
- Vorgehen: Aufteilung des Netzwerks in zwei Ströme zur Schätzung des Zustandswerts und der Vorteile. Kombination beider zur Schätzung der Q-Werte
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem Aktionsraum
- Beispielhafte Gym-Umgebung: „Pong-v0“

2015: Prioritized Experience Replay (PER):

- Ziel: Steigerung der Lerneffizienz durch priorisiertes Sampling von Erfahrungen
- Vorgehen: Priorisieren der Erfahrungen basierend auf dem TD-Fehler für das Replay
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem Aktionsraum
- Beispielhafte Gym-Umgebung: „MountainCarContinous-v0“

2015: Deep Deterministic Policy Gradient (DDPG):

- Ziel: Anwendung von Deep Learning auf Fragestellungen mit kontinuierlichem Aktionsraum
- Vorgehen: Verwendung einer Strategie und Wertefunktion, die durch tiefe Netzwerke repräsentiert werden, und Optimierung dieser durch das Policy-Gradient und Q-Learning
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem Aktionsraum
- Beispielhafte Gym-Umgebung: „Pendulum-v0“

2016: Asynchronous Advantage Actor-Critic (A3C):

- Ziel: Verbesserung der Stabilität und Effizienz von Deep Learning durch asynchrone Updates
- Vorgehen: Verwendung von mehreren parallelen Agenten, die jeweils unabhängig voneinander lernen und ihre Parameter asynchron aktualisieren
- Einsatzbereiche: Vielseitig, insbesondere Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem oder kontinuierliche Aktionsraum
- Beispielhafte Gym-Umgebung: „SpaceInvaders-v0“

2017: Proximal Policy Optimazation (PPO):

- Ziel: Verbesserung der Stabilität und Effizienz von Policy Gradient Methoden
- Vorgehen: Begrenzung der Änderungen der Policy bei jedem Update, um starke Abweichungen zu vermeiden
- Einsatzbereiche: Vielseitig, vor allem Fragestellungen mit hohen dimensionalem Zustandsraum und diskretem oder kontinuierliche Aktionsraum
- Beispielhafte Gym-Umgebung:

„Highway-v0“

2018: Soft Actor-Critic (SAC):

- Ziel: Maximierung der Entropie neben der erwarteten Belohnung
- Vorgehen: Nutzung einer stochastischen Policy und Optimieren des Trade-off zwischen erwarteter Belohnung und Entropie
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und kontinuierlichem Aktionsraum
- Beispielhafte Gym-Umgebung: „HalfCheetah-v2“

2018: Twin Delayed DDPG (TD3):

- Ziel: Verbesserung der Stabilität von DDPGs durch Adressierung der Funktionsschätzungsfehler
- Vorgehen: Nutzung von zwei Q-Funktionen und Verzögerung der Aktualisierung der Policy
- Einsatzbereiche: Fragestellungen mit hohen dimensionalem Zustandsraum und kontinuierlichem Aktionsraum
- Beispielhafte Gym-Umgebung: „Ant-v2“