

차 례

차 례	1
제 1 장 강화학습 문제	3
1.1 강화학습	3
1.2 사례	6
1.3 강화학습의 요소	7
1.4 한계와 범위	9
1.5 틱택토 사례 탐구	11
1.6 요약	15
1.7 강화학습의 초기 역사	16
제 I 편 Tabular Solution Methods	25
제 2 장 Multi-armed Bandits	26
제 3 장 Finite Markov Decision Processes	27
제 4 장 Dynamic Programming	28
제 5 장 몬테카를로 기법	29
제 6 장 Temporal-Difference Learning	31
제 7 장 Multi-step Bootstrapping	32
제 8 장 Planning and Learning with Tabular Methods	33
제 II 편 Approximate Solution Methods	34
제 9 장 On-policy Prediction with Approximation	35

9.1 Value-function Approximation	35
9.2 The Prediction Objective (MSVE)	35
9.3 Stochastic-gradient and Semi-gradient Methods	35
9.4 Linear Methods	35
9.5 Feature Construction for Linear Methods	35
9.6 Nonlinear Function Approximation: Artificial Neural Networks	35
제 10 장 On-policy Control with Approximation	36
제 11 장 Off-policy Methods with Approximation	37
제 12 장 Eligibility Traces	38
제 13 장 Policy Gradient Methods	39
제 III 편 Looking Deeper	40
제 14 장 Psychology	41
제 15 장 Neuroscience	42
제 16 장 Applications and Case Studies	43
참고 문헌	44

제 1 장

강화학습 문제

학습에 대해 생각할 때 아마도 가장 먼저 떠오르는 생각은 우리가 환경과 상호작용하며 배운다는 것이다. 갓난아이가 팔을 흔들며 둘러볼 때 어떻게 하라고 알려주는 사람이 없지만, 유아는 감각운동적으로 환경과 직접 연결돼있다. 감각운동을 연습하면 인과 관계와 행동의 결과 그리고 목적을 달성하기 위한 행동 순서에 대한 풍부한 정보가 쌓인다. 살면서 이런 상호작용을 통해 환경과 우리 자신에 대한 대부분의 지식을 얻는다. 운전이나 대화하는 방법을 배울 때 우리는 환경이 우리 행동에 어떻게 반응하는지 민감하게 인식하며 앞으로 일어날 일에 영향을 주려고 행동한다. 상호작용을 통한 학습이 거의 모든 학습과 지능 이론의 근간이 된다.

이 책은 상호작용을 통한 학습을 계산적으로 접근하는 방법을 탐구한다. 사람과 동물이 어떻게 학습하는지를 직접 이론화하기보다는 이상적인 학습 환경을 살펴보고 다양한 학습 기법의 효율성을 평가한다. 즉, 우리는 인공지능 연구자나 공학자의 관점으로 본다. 과학적 혹은 경제적으로 가치가 있는 학습 문제를 효율적으로 푸는 기계를 어떻게 설계할지 탐구한다. 각각의 설계를 수학적으로 분석하거나 컴퓨터로 실행하여 검토해본다. 강화학습(*reinforcement learning*)이라는 우리의 접근 방식은 다른 기계학습 보다 상호작용을 통한 훨씬 더 목표 지향적인 학습에 집중한다.

1.1 강화학습

기계학습(machine learning)과 등산(mountaineering)과 같이 이름이 “ing”로 끝나는 강화학습은 해결할 문제이자 동시에 이런 문제에 효과적인 해결책이며 이런 문제와 해결 방법을 연구하는 분야이기도 한다. 강화학습 문제는 숫자로 된 보상 신호를 최대로 늘리기 위해 무엇을 할지 (어떻게 상황을 행동으로 연결할지) 배우는 것이다. 학습 시스템의 행동이 이후 자신의 입력에 영향을 주기 때문에 본질적으로 폐회로(*closed-loop*) 문제이다. 게다가 다른 기계학습 형태와 달리 학습자는 무슨 행동을 할지 지시받지 않는다. 대신 직접 시도해가며 어떤 행동이

가장 많은 보상을 주는지 발견해야 한다. 흥미롭고 도전적인 대부분의 사례에서 행동은 당장의 보상뿐 아니라 다음 상황에 영향을 미쳐서 이후 모든 보상을 좌우한다. ① 본질적으로 폐회로 ② 어떤 행동을 할지 직접 지시를 받지 않음 ③ 보상 신호 등 행동의 장기간 영향, 이 세 가지가 강화학습 문제의 가장 중요한 특징이다.

3장에 가서야 강화학습 문제를 최적 제어 분야의 마르코프 결정 프로세스 (Markov decision process, MDP)로 표현하지만, 이 기본 발상은 목표를 달성하기 위해 환경과 상호작용하는 학습 에이전트의 실제 문제에서 가장 중요한 측면이다. 분명히 이런 에이전트는 어느 정도 환경 상태를 파악할 수 있고, 상태에 영향을 주는 행동을 취해야 한다. MDP 공식은 단지 감각과 행동 그리고 목표라는 세 가지 측면을 지나치게 단순화하지 않는 선에서 가장 간단히 표현하려고 한다. 이런 문제를 푸는데 적합한 기법이라면 무엇이든지 강화학습 기법이라고 한다.

강화학습은 현재 가장 활발하게 연구되는 기계학습 분야인 지도학습 (*supervised learning*)과 다르다. 강화학습은 지식을 가진 외부 감독자가 라벨을 붙인 예제로 구성된 훈련 세트를 가지고 학습한다. 예제는 상황 설명과 함께 (보통은 상황이 어떤 분류에 속하는지 파악하기 위해) 그 상황에서 시스템의 올바른 행동 (라벨)을 지시한다. 이런 학습의 목표는 반응을 추론하고 일반화하여 훈련 세트에 없는 상황에서도 시스템이 올바로 동작하게 만드는 것이다. 중요한 학습 분야이지만, 이것만으로는 상호작용으로 배우는데 충분하지 않다. 보통은 상호작용 문제에서 올바르고 에이전트가 만나는 모든 상황을 대표하는 바람직한 행동을 알기 어렵다. 학습이 절실한 미지의 영역에서 에이전트는 자신의 경험을 통해 배울 수 있어야 한다.

또한 강화학습은 기계학습 연구자들이 비지도학습 (*unsupervised learning*)이라고 부르는 것과도 다르다. 대개 비지도학습은 라벨이 없는 자료 묶음에 숨겨진 구조를 발견하려 한다. 기계학습 분야를 지도학습과 비지도학습, 두 개만으로 나눌 수 있는 것처럼 보이지만 그렇지 않다. 올바른 행동 예제가 없어도 되기 때문에 강화학습을 비지도학습의 일종으로 보고 싶을지도 모른다. 그러나 강화학습은 숨은 구조를 찾기보다는 보상 신호를 최대로 늘리려고 노력한다. 에이전트의 경험에 숨겨진 구조를 발견하는 것도 강화학습에 도움이 되지만, 그것만으로는 보상 신호를 최대화하는 강화학습 에이전트 문제를 해결할 수 없다. 그러므로 우리는 지도학습과 비지도학습 등과 더불어 강화학습을 기계학습의 세 번째 분야로 본다.

다른 종류의 학습과 달리 강화학습에서 두드러지는 어려움 중 하나는 탐색과 활용 사이의 균형이다. 많은 보상을 얻으려면, 강화학습 에이전트는 과거에 시도하여 효율적으로 보상을 준다고 알려진 행동을 선호해야 한다. 그러나 효율적으로 보상을 주는 행동을 찾으려면 과거에 선택하지 않은 행동을 시도해보아야 한다. 에이전트는 보상을 받으려면 이미 알고 있는 것을 활용해야 (*exploit*) 하지만, 또한

미래에 더 나은 선택을 하기 위해 탐색해야(explore) 한다. 작업에 성공하려면 집중과 탐색 중 하나만 전적으로 추구하면 안 되는 점이 모순이다. 에이전트는 다양한 행동을 시도하면서 동시에 최고로 보이는 행동을 점진적으로 선호해야 한다. 통계적 작업이라면, 에이전트는 보상의 기댓값을 신뢰할 수 있게 예측하기 위해 여러 번 시도해야 한다. 탐색-활용 모순은 수학자들이 수십 년 동안 열심히 연구했다(2장 참고). 일단 탐색과 활용의 균형 문제는 최소한 순수한 형태의 지도학습과 비지도학습에는 없다고 말할 수 있다.

강화학습의 다른 특징은 불확실한 환경과 상호작용하는 목적 지향적 에이전트의 문제 전체를 분명히 고려하는 점이다. 큰 그림에 어떻게 부합하는지 말하지 않은 채 문제의 부분들을 고려하는 여러 접근 방식과 다르다. 예를 들어, 지도학습과 연관된 상당수의 기계학습 연구는 결국 그 능력을 어떻게 유용하게 사용할지 말하지 않는다. 범용 목표를 향한 계획 이론을 개발하는 다른 연구자들은 실시간 의사결정시 계획의 역할이나 계획에 필요한 예측 모델을 어떻게 만들지 고려하지 않는다. 이런 접근 방법도 여러 유용한 결과를 달성했지만, 문제 일부분만 격리하여 고려한 점이 큰 한계이다.

반대로 강화학습은 완전하고 상호작용하는 목적 지향 에이전트에서 시작한다. 모든 강화학습 에이전트는 명확한 목표를 가지고, 주변환경을 인지하고, 환경에 영향을 주는 행동을 선택할 수 있다. 게다가 강화학습은 처음부터 항상 매우 불확실한 주변환경에도 불구하고 에이전트는 동작해야 한다고 가정한다. 강화학습에 계획이 들어가면, 계획과 실시간 행동 선택 사이의 상호 작용과 어떻게 환경 모델을 얻고 향상할지 다루어야 한다. 강화학습에 지도학습이 들어간다면, 어떤 능력이 결정적인지 판단해야 할 이유 때문이다. 학습 연구를 진행하기 위해 중요한 문제를 부분으로 나누어 연구해야 했다. 그러나 아직 완전한 에이전트의 세부 요건을 파악할 수 없지만, 부분 문제는 완전하고 상호작용하는 목적 지향 에이전트에서 명확할 역할이 있어야 한다.

여기서 완전하고 상호작용하는 목적 지향 에이전트란 완전한 유기체나 로봇 같은 것을 뜻하지 않는다. 예를 들면, 완전하고 상호작용하는 목적 지향 에이전트는 동작하는 더 큰 시스템의 구성요소일 수도 있다. 이 경우 에이전트는 더 큰 나머지 시스템과 직접 상호작용하며 더 큰 시스템의 환경과 간접적으로 상호작용 한다. 앞으로 로봇의 배터리 충전 수준을 살피고 로봇의 제어 아키텍처에 명령을 보내는 에이전트를 간단한 예로 들 것이다. 이 에이전트의 환경은 로봇의 환경과 로봇의 나머지 부분이다. 강화학습 프레임워크의 일반성을 평가하려면 가장 뼈한 에이전트와 환경 너머를 보아야 한다.

현대 강화학습의 가장 흥미로운 점 중 하나는 다른 공학과 과학 분야와 상당히 교류한 결실이다. 강화학습은 십여 년간 인공지능과 기계학습이 통계학과 최적화 그리고 다른 수학 분야와 넓게 통합하는 추세 한가운데에 있다. 예를 들어, 파라미터 근사기(parameterized approximator)를 가지고 학습하는 일부 강화학습

기법은 운용 과학(operations research)과 제어 이론의 고전적인 “차원의 저주”를 다룬다. 또 강화학습은 심리학과 신경과학과 밀접하게 교류하며 서로에게 커다란 기여를 했다. 기계학습 분야 중에 강화학습은 인간과 다른 동물의 학습과 가장 가깝고, 강화학습의 여러 핵심 알고리즘은 원래 생물학적 학습 시스템에서 유래했다. 그리고 강화학습은 반대로 실험 결과에 더 부합하는 동물 학습의 심리적 모델과 뇌의 보상 시스템 일부에 대한 유력한 모델을 기여했다. 이 책은 공학과 인공지능에 관련된 강화학습 아이디어와 함께 14장과 15장에서 심리학과 신경과학과 연관성을 다룬다.

마지막으로 강화학습은 단순한 일반 이론을 향한 인공지능의 큰 추세의 일부 이기도 하다. 1960년대 말 이후 많은 인공지능 연구자들은 일반 이론이란 없으며 지능은 대신 특수한 목적을 가지는 재주와 절차와 휴리스틱이 엄청나게 모인 결과라고 짐작했다. 수백만 혹은 수십억 개의 충분한 사실을 기계에 넣으면 지능이 생길 것이라는 말도 종종 있었다. 검색과 학습 같은 일반 이론에 바탕을 둔 방법은 “약한 방법(weak methods)”으로 분류하고, 특정 지식에 근거한 방법은 “강한 방법(strong methods)”이라고 불렸다. 이런 시각은 여전히 흔하지만, 훨씬 덜 지배적이다. 우리가 보기에도 그런 시각은 시기상조다. 일반 이론이 없다고 결론짓기에는 연구가 너무 부족했다. 현대 인공지능은 방대한 도메인 지식을 반영하려는 노력과 더불어 학습과 검색 그리고 의사결정의 일반 이론을 찾는 연구를 많이 한다. 시계추가 얼마나 돌아왔는지 확실하지 않지만, 강화학습 연구는 단순하고 적은 수의 인공지능 일반 이론을 추구한다.

1.2 사례

강화학습의 발전을 이끈 일부 사례와 가능한 적용사례를 보면 강화학습을 이해하기 좋다.

- 체스 고수가 말을 움직인다. 상대방의 수와 그 다음 수를 예상한 계획과 특정 위치와 움직임이 좋은지 순간적인 직관을 통해 어떤 말을 움직일지 선택한다.
- 적응 제어기는 석유 정제소 동작 파라미터를 실시간으로 조절한다. 제어기는 기술자가 처음에 짐작한 값을 무조건 따르지 않고 지정한 한계 비용을 토대로 생산량/비용/품질 사이의 균형을 최적화한다.
- 새끼 가젤은 태어나고 몇 분간 일어서려고 애쓴다. 반 시간이 지나면 시간당 30킬로미터 속력으로 달린다.
- 이동 로봇은 쓰레기를 주으려고 새로운 방에 들어갈지 아니면 배터리 충전 소로 되돌아가 시작해야 할지 결정한다. 현재 배터리 충전 상태와 과거에 얼마나 빠르고 손쉽게 충전소를 찾았는지를 가지고 결정을 내린다.

- 필이 아침 식사를 준비한다. 찬장으로 걸어가기, 찬장 열기, 시리얼 상자 선택하기, 상자로 팔을 뻗어서 상자를 잡고 가져오기. 이렇게 유심히 살펴보면 일상적인 활동조차도 조건부 행동과 서로 맞물린 목표들 사이의 관계들이 복잡하게 엮여있다. 그릇과 숟가락 그리고 우유병을 잡기 위해서도 일련의 복잡하고 조화된 상호작용 동작이 필요하다. 단계별로 정보를 습득하고 손과 운동을 안내하기 위해 눈을 움직여야 한다. 어떻게 물건을 옮기고 무엇을 먼저 선택으로 가져오면 좋을지를 끊임없이 빠르게 판단한다. 숟가락 잡기와 냉장고로 이동하기 같은 목표가 매 단계를 안내하고, 이 목표들은 숟가락으로 이미 준비한 시리얼을 먹어서 궁극에는 영양 섭취 같은 다른 목표에 기여한다. 인지하든 모르든 관계없이 필은 자신의 몸 상태 정보를 가지고 필요한 영양분과 배고픈 정도 그리고 선호하는 음식을 파악한다.

이 사례들은 너무 기본적이어서 간과하기 쉬운 공통된 특징이 있다. 모두 능동적인 행동 결정 에이전트와 환경 사이의 상호작용이 있다. 환경의 불확실성에도 불구하고 에이전트는 목표를 달성할 방법을 찾는다. 에이전트의 행동은, 예를 들어, 다음 체스 말 위치, 정제소 비축량, 로봇의 다음 위치와 미래 배터리 충전 수준 등, 환경의 미래 상태에 영향을 주고, 그것이 다시 에이전트의 미래 선택지와 기회를 바꿀 수 있다. 올바른 선택을 하려면 행동의 간접적이고 지연된 결과를 고려해야 하고, 그래서 전망이나 계획이 필요할지 모른다.

또한, 이런 사례에서 행동의 영향을 완전히 예상할 수 없다. 그래서 에이전트는 환경을 자주 감시하고 적절하게 반응해야 한다. 예를 들어, 필은 우유를 시리얼 그릇에 부을 때 넘치지 않게 지켜봐야 한다. 모든 사례에서 에이전트는 직접 감각한 것을 가지고 목표까지 얼마나 진행했는지 판단할 수 있는 점에서 목표가 명시적이다. 체스 선수는 자신의 승패를 알고, 정제소 제어기는 석유 생산량을 알고, 이동 로봇은 언제 배터리가 바닥날지 알고, 필은 자신이 아침 식사를 즐기는지 안다.

모든 사례에서 에이전트는 경험을 통해 시간이 지나면서 나아진다. 체스 선수는 자리를 보는 통찰을 갖고닦아 결국 경기를 향상한다. 새끼 가젤은 달리기 효율을 향상한다. 필은 자연스럽게 아침 식사를 준비하는 방법을 배운다. 에이전트가 작업을 시작할 때 가진 (과거 비슷한 작업 경험 혹은 설계나 진화로 알게 된) 지식이 학습에 도움을 주지만, 작업 특성에 맞추어 행동을 조정하기 위해 환경과 상호작용은 꼭 필요하다.

1.3 강화학습의 요소

에이전트와 환경 이외에 강화학습 시스템의 주된 요소는 정책(*policy*), 보상 신호(*reward signal*), 가치함수(*value function*) 그리고 선택적인 환경 모델(*model*), 이렇게 네가지다.

정책(*policy*)은 어떤 시점에서 학습 에이전트의 행동 방식을 정의한다. 간단히 말해서 정책은 인지된 환경 상태를 그 상태에서 취할 행동으로 대응한다. 심리학에서 (동물 내재 자극을 포함한) 일련의 자극-반응(stimulus-response) 규칙 혹은 연상(association)에 해당한다. 정책이 간단한 함수 혹은 참조표인 경우도 있고, 검색 절차같이 엄청난 계산이 필요한 정책도 있다. 정책만으로 행동을 결정하기 충분한 점에서 정책은 강화학습 에이전트의 핵심이다. 일반적으로 확률적인 정책도 가능하다.

보상 신호(*reward signal*)는 강화학습 문제의 목표를 정의한다. 매시간 단계마다 환경은 하나의 숫자 즉 보상을 강화학습 에이전트에게 보낸다. 장기적인 보상 총합을 최대화하는 것이 에이전트의 유일한 목표다. 그래서 보상 신호는 에이전트에게 무엇이 좋고 무엇이 나쁜지 정의한다. 보상을 생물학적 시스템에서 즐거움이나 고통을 경험하는 것으로 생각할 수 있다. 보상은 즉시 발생하고, 에이전트가 마주한 문제를 정의하는 특징이다. 그러므로 에이전트는 보상 신호를 생성하는 프로세스를 변경할 수 없다. 에이전트는 직접 자신의 행동을 통해 프로세스가 발생하는 보상을 변경할 수 있다. 보상 신호가 환경 상태에 의존하기 때문에 에이전트는 환경 상태를 변경하여 간접적으로 보상에 영향을 줄 수도 있다. 그러나 에이전트는 보상을 만드는 함수를 변경할 수는 없다. 즉, 에이전트는 자신이 대면한 문제를 다른 문제로 바꿀 수 없다. 보상 신호는 보상을 변경하는 기본 근거이다. 정책으로 선택한 행동이 낮은 보상을 낸다면, 미래에 같은 상황에서 다른 상태를 선택하도록 정책을 변경할 수 있다. 일반적으로 보상 신호는 환경 상태와 선택한 행동의 확률함수이다. 3장에서 에이전트가 보상 함수를 변경할 수 없다는 개념이 생물학에서 동물 뇌에 생성된 보상 신호를 관찰한 결과와 어떻게 일치하는지 설명한다.

보상 신호가 당장에 무엇이 좋은지 알려준다면, 가치함수(*value function*)는 장기적으로 무엇이 좋은지 알려준다. 쉽게 말해서 상태의 가치는 에이전트가 그 상태부터 시작하여 앞으로 누적해서 받길 기대하는 보상의 총합이다. 보상이 환경 상태에 내재한 즉각적인 바람직함이라면, 가치는 따라갈 가능성성이 높은 상태와 그 상태들의 보상들을 고려한 장기적인 바람직함이다. 예를 들어, 당장 보상은 적지만, 그다음 상태들의 보상이 커서 가치가 높은 상태가 있을 수 있고, 반대도 가능하다. 사람에 비유하면 보상은 (높다면) 기쁨이나 (낮다면) 고통과 비슷하다면, 가치는 어떤 환경 상태에서 우리가 얼마나 즐겁고 불쾌한지를 더 원시안적으로 가늠한 것이다. 이제 가치 함수 개념에 익숙하길 바란다.

어떤 의미에서 보상은 기본적이고, 보상을 예측하는 가치는 부수적이다. 보상 없이는 가치란 있을 수 없으며 가치를 예측하는 유일한 이유는 더 많은 보상을 받기 위해서다. 그럼에도 불구하고 우리가 결정을 내리고 결정을 평가할 때 가치를 주로 고려한다. 우리는 가장 높은 보상이 아니라 가장 높은 가치를 주는 상태로 가는 행동을 추구한다. 이런 행동이 장기적으로 가장 큰 보상을 주기

때문이다. 불행하게도 보상을 측정하는 것보다 가치를 측정하기가 훨씬 어렵다. 보상은 기본적으로 환경이 직접 주지만, 가치는 짐작해야 하고 에이전트가 일생 동안 관찰한 내용을 가지고 계속 다시 가치를 추정해야 한다. 사실 효율적으로 가치를 예측하는 방법 우리가 고려할 거의 모든 강화학습 알고리즘에서 가장 중요한 요소이다. 분명히 가치 예측은 지난 수십 년 동안 강화학습 연구에서 가장 중요한 역할을 차지했다.

마지막 네 번째 강화학습 시스템 요소는 환경 모델(*model*)이다. 모델은 환경의 행동을 흉내내고, 일반적으로 환경이 어떻게 동작할지 추론할 수 있게 해준다. 예를 들어, 상태와 행동이 주어지면, 모델은 다음 상태와 다음 보상을 예측할 수 있다. 계획(*planning*)에도 모델을 사용한다. 계획은 실제 경험하지 않고 가능한 미래 상황을 고려하여 일련의 행동을 결정하는 방법을 말한다. 모델과 계획을 사용하여 강화학습 문제를 푸는 기법을 모델 기반(*model-based*) 기법이라고 하고, 계획의 거의 반대인 시행착오(*trial-and-error*)를 통해 학습하는 단순한 비모델(*model-free*) 기법도 있다. 8장에서 우리는 시행착오학습과 함께 환경 모델을 학습하고 모델을 사용하여 계획하는 강화학습 시스템을 살펴볼 것이다. 현대 강화학습은 저수준의 시행착오 학습부터 고수준의 신중한 계획까지 모두 망라한다.

1.4 한계와 범위

이 책에서 다루는 대부분의 강화학습 기법은 가치함수 예측과 관련이 있지만, 강화학습 문제를 풀기 위해 반드시 가치함수를 추정할 필요는 없다. 예를 들어, 그동안 가치함수 없이도 유전 알고리즘, 유전 프로그래밍, 시뮬레이티드 어닐링(simulated annealing) 등 다른 최적화 기법을 사용하여 강화학습 문제를 풀었다. 이 기법들은 서로 다른 정책을 사용하여 환경과 상호작용하는 비학습 에이전트들의 “평생” 행동을 평가하고, 가장 많은 보상을 받은 에이전트를 뽑는다. 개별 삶 동안 계체들이 학습하지 않고도 가장 능숙한 행동을 한 계체를 만드는 생물학적 진화와 유사하기 때문에 진화(*evolutionary*) 기법이라고 부른다. 정책 집합의 크기가 충분히 작거나 좋은 정책이 흔하거나 찾기 쉬운 구조라면 (혹은 충분히 많은 시간 동안 탐색할 수 있다면), 진화 기법이 유용하다. 또한, 진화 기법은 학습 에이전트가 환경 상태를 정확히 감지할 수 없는 문제에 유리하다.

우리는 환경과 상호작용하면서 학습하는 강화학습 기법에 집중한다. (몇몇 연구에서처럼 학습 알고리즘을 진화하지 않는 한) 진화 기법은 그렇지 않다. 우리는 많은 경우 자세한 개별 상호작용 행동을 활용하는 기법이 진화 기법보다 훨씬 더 효율적이라고 믿는다. 진화 기법은 강화학습 문제의 여러 유용한 구조를 무시한다. 즉, 진화 기법이 찾는 정책이 상태를 행동으로 변환하는 함수라는 사실을 사용하지 않고, 개체가 평생 동안 경험한 상태와 행동을 무시한다. 상태를 오해한 경우처럼 이런 정보가 오히려 혼란을 주는 경우가 없진 않지만, 더 효율적인 검색이 가능한 경우가 많다. 진화와 학습이 많은 점을 공유하고 서로 자연스럽게

협력하지만, 우리는 진화 기법 자체가 강화학습 문제에 특별히 적합하다고 보지 않는다. 단언하면 이 책에 나오는 “강화학습 기법”에는 진화 기법이 없다.

그러나 진화 기법같이 가치함수를 사용하지 않는 기법을 몇 가지 다룬다. 수치파라미터들로 정의된 정책 집합을 검색하는 기법들이다. 정책의 성능을 가장 빨리 향상하기 위해 파라미터를 어떤 방향으로 조정할지를 추정한다. 그러나 진화 기법과 달리 에이전트가 환경과 상호작용하며 예측하기 때문에 개별 상호작용 행동 내역을 활용할 수 있다. 정책 기울기 기법(*policy gradient method*)이라고 부르는 이런 기법은 여러 문제에서 유용했고, 일부 가장 간단한 강화학습 기법들이 여기에 속한다. 사실 일부는 기울기를 더 잘 예측하려고 가치함수를 추정하기도 한다. 크게 보면 정책 기울기 기법과 우리가 생각하는 강화학습 기법은 명확하게 나뉘지 않는다.

흔히 오해하기 때문에 강화학습과 최적화 기법의 연관성은 더 언급한다. 수치보상 신호를 최대화하는 것이 강화학습 에이전트의 목표라고 말할 때 에이전트가 최대 보상을 달성해야 한다는 의미가 아니다. 최대를 만들려는 시도가 최대가 되었다는 뜻은 아니다. 요점은 강화학습 에이전트가 항상 자신이 받는 보상을 최대로 늘리려고 시도한다는 점이다. 보상의 최댓값이 존재하더라도 다양한 이유 때문에 최대 보상을 얻지 못할 수 있다. 즉, 최적화는 최적이 아니다.

최적화 기법이 최적을 달성하는지와 관계없이 최적화에 기반한 인공지능 시스템은 시스템의 행동을 항상 예상할 수 없기 때문에 주의해야 한다. 강화학습 에이전트는 종종 환경이 보상을 주는 예상하지 못한 경우를 찾는다. 어떤 면에서 일종의 창의력으로 지능의 바람직한 속성이다. 진화와 강화학습의 정수인 변형과 선택에 의한 프로세스는 동물이나 인공지능이 어떤 도전에 직면하더라도 새로운 성공 방식을 발견한다. 그러나 어떻게 이 예상 못 한 “해결책”이 의도하지 않거나 바람직하지 않은 결과를 가져오지 않을지 확신할 수 있을까?

이런 걱정은 강화학습에서 낯설지 않다. 고테의 1797년 시 “마법사의 제자” 같이 문학에서 주된 주제였고, “무언가를 바랄 때는 실제로 생길 수 있으니 조심해라”라는 말로 요약할 수 있다. 최적화를 할 때 제한을 강제하거나 위험에 민감하도록 최적화 목적 함수를 조정하는 등 이 문제의 심각성을 줄이려는 시도는 완전한 해결책이 아니다. 표준 공학 절차는 사람들의 안전하게 여기는 제품이나 구조물 그리고 실제 시스템을 만들기 전에 최적화 과정의 결과를 면밀하게 조사해 왔다. 이점은 강화학습을 공학적으로 사용할 때도 필수적이고, 강화학습 시스템을 (강화학습 에이전트는 물론이고 에이전트의 환경과 그 환경 안에 있는 사람에게) 예기치 않은 결과를 용납할 수 없는 분야에 적용할 때 특히 매우 조심해야 한다. 인공지능이 빠르게 발전하여, 특히 기계학습 시스템이 초인적인 성능을 내는 특정 분야에서 이런 우려가 현실이 된다.

최적화의 비예측성은 강화학습 시스템을 실세상에 책임 있게 적용하는 더 큰 주제의 일부일 뿐이다. 여기서도 다른 공학 기술에 대한 걱정과 크게 다르지

않고, 원하지 않는 결과가 생길 위험을 줄이는 다양한 방식이 개발되었다. 특히 최적 제어 기법의 위험을 줄이는 방법이 강화학습에 적용되었다. 크고 복잡한 주제이기 때문에 입문서가 다룰 범위를 넘어선다. 그러나 단지 학습과 지능에 대한 이론이 아니라 공학적 방법론으로 볼 때 강화학습도 두말할 나위 없이 어느 공학적 방법론과 마찬가지로 주의를 기울여야 한다.

1.5 틱택토 사례 탐구

강화학습의 일반 개념을 설명하고 다른 방식과 비교하기 위해 한 가지 사례를 더 자세히 살펴보자.

어린 시절 친숙한 틱택토(tic-tac-toe) 게임을 보자.

두 선수는 가로세로 세 칸인 보드에 번갈아 말을 둔다. 한 선수는 X를 다른 선수는 O를 내고, 오른쪽 그림에서 X 선수같이 한쪽이 가로나 세로 혹은 대각선 방향으로 세 칸을 두면 이긴다. 어느 선수도 세 개를 연달아 두지 못하고 칸이 모두 차면 비긴다. 잘 하는 선수는 절대로 안 지도록 말을 둘 수 있기 때문에 우리가 이길 수 있도록 종종 실수하는 불완전한 상대 선수를 가정한다. 여기서는 지거나 비기면 똑같이 안 좋다고 보자. 상대의 약점을 찾고 이길 가능성을 최대화하도록 배우는 선수를 어떻게 만들까?

X	O	O
O	X	X
		X

간단한 문제이지만, 고전적인 방식으로는 만족스럽게 해결하기 쉽지 않다. 예를 들어, 게임이론의 고전적인 “미니맥스(minimax)” 방식은 상대가 특이한 식으로 게임을 한다고 가정했기 때문에 여기에 알맞지 않다. 미니맥스 선수라면 조금이라도 질 수 있는 수라면 미숙한 상대 때문에 항상 이길 수 있을지라도 절대로 두지 않는다. 동적 프로그래밍(dynamic programming) 같은 연속 결정 문제를 위한 전통적인 최적화 기법은 어떤 상대에 대한 최적해를 계산할 수 있지만, 모든 보드 상태에 상대가 둘 수 확률을 포함하여 상대에 대한 완전한 명세를 입력으로 주어야 한다. 실용적으로 관심이 있는 대부분의 문제와 마찬가지로 이런 정보를 미리 주지 않는다고 가정한다. 한편 상대와 여러 번 게임을 하여 경험을 통해 상대의 정보를 추측할 수 있다. 이 문제에서 아마도 가능한 최선의 방법은 먼저 어느 정도 신뢰할 수 있는 수준까지 상대방 행동 모델을 학습하고 상대의 근사 모델을 가지고 동적 프로그래밍을 적용하여 최적해를 계산하는 것이다. 결국, 이 책에서 다룰 강화학습 기법과 크게 다르지 않다.

이 문제에 진화 기법을 적용하면, 가능한 정책 집합에서 상대방을 이길 확률이 높은 정책을 직접 검색한다. 여기서 정책은 가로세로 세 칸인 보드 위에 X와 O를 조합한 모든 게임 상태마다 선수가 어디에 말을 둘지 알려주는 규칙이다. 정책마다 상대와 일정 횟수 게임을 하여 이길 확률을 추측할 수 있다. 그런 다음 앞으로 무슨 정책을 사용할지 정한다. 전형적인 진화 기법은 계속 정책을 만들

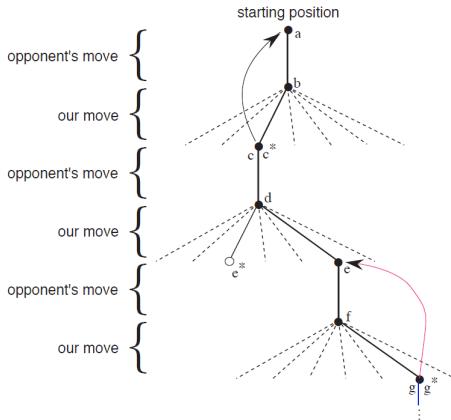


그림 1.1: 일련의 틱택토 움직임. 게임에서 선택한 움직임은 실선이고, 우리(강화학습 선수)가 고려했지만 선택하지 않은 움직임은 점선으로 표시한다. 두 번째 움직임은 더 우선순위가 높은 e^* 로 이어진 길이 옆에 있지만 대신 선택한 탐색적 움직임이다. 탐색적 움직임은 학습하지 않지만, 다른 모든 움직임은 곡선 화살표로 보이는 백업으로 학습한다. 백업은 본문을 참고하라.

고 평가하여 정책 집합에서 정책을 점진적으로 향상한다. 혹은 정책 개체군을 유지하고 평가하는 유전 방식 알고리즘을 사용할 수 있다. 말 그대로 수백 가지 최적화 기법이 가능하다.

가치함수를 사용하여 어떻게 틱택토 문제를 푸는지 보자. 먼저 게임의 모든 상태마다 숫자를 붙인 표를 만든다. 숫자는 가장 최근에 추측한 그 상태에서 승리할 확률이다. 이 추측값을 상태의 값으로, 전체 표는 학습한 가치함수로 본다. 현재 A에서 시작하여 승리할 확률이 B에서 시작하는 경우보다 높다고 예측했다면, 상태 A가 상태 B 보다 높은 값을 가진다 혹은 더 “좋다고” 말한다. 우리가 항상 X를 둔다면, X가 세 개 줄지어 있는 모든 상태의 승리 확률은 이미 승리했기 때문에 1이다. 비슷한 식으로 O를 세 개 “채운” 모든 상태는 우리가 그 상태에서 이길 수 없기 때문에 이길 확률이 0이다. 다른 모든 상태의 초기값은 이길 확률을 50%로 짐작하여 0.5로 잡는다.

상대와 여러 번 게임을 한다. 칸을 선택할 때 (비어있는) 둘 수 있는 칸들을 하나씩 채운 상태를 구하고, 표에서 상태들의 현재 값을 찾는다. 대부분의 경우 탐욕적으로(*greedily*) (다른 말로 이길 가능성성이 제일 높은) 가장 큰 값을 가진 상태로 이동하는 행동을 선택한다. 그러나 가끔은 무작위로 행동을 선택하기도 한다. 무작위로 선택하지 않으면 절대로 만나지 못할 상태를 경험하기 때문에 탐색적(*exploratory*) 이동이라고 부른다. 게임 중에 고려하고 이동한 움직임들을 그림 1.1과 같이 그릴 수 있다.

게임을 하면서 경험을 통해 상태 값을 변경한다. 그래서 더 정확하게 승리할 확률을 예측하려고 한다. 그림 1.1의 화살표처럼 탐색적 움직임을 할 때마다 이동 후 상태 값을 이전 상태로 “백업(backup)”한다. 정확히는 이전 상태의 현재 값을

이후 상태의 값에 가깝게 조정한다. 이전 상태의 값을 이후 상태의 값 방향으로 이전 상태의 값을 일정 비율만큼 이동하면 된다. 탐욕적 움직임 이전 상태를 s , 이후 상태를 s' 라고 쓰면, s 의 예측값 $V(s)$ 는 다음과 같다.

$$V(s) \leftarrow V(s) + \alpha[V(s') - V(s)]$$

여기서 α 는 학습 속도를 조절하는 스텝 크기 파라미터(*step-size parameter*)라는 작은 양의 분수이다. 이 개신 규칙은 시간차(*temporal-difference*) 학습 기법의 일종이다. 시간차란 이름은 두 다른 시점의 예측 차이 $V(s') - V(s)$ 를 기초로 변경하기 때문이다.

위에서 말한 기법은 이 작업에 상당히 적합하다. 예를 들어, 시간이 지나면서 스텝 크기 파라미터를 적절하게 줄인다면(26쪽 참고), 이 기법은 우리 선수가 모든 상태에서 시작하여 최적의 플레이를 통해 어떤 고정된 상태를 이길 정확한 확률로 수렴한다. 게다가 (탐색적 움직임을 제외하고) 선택한 움직임은 사실 (불완전한) 상대에 대한 최적의 움직임이다. 즉, 상대와 경기하는 최적의 정책으로 수렴한다. 스텝 크기 파라미터가 점차 영으로 줄어들지 않는다면, 이 선수는 자신의 게임 방식을 천천히 바꾸는 선수를 상대로도 잘 싸운다.

이 사례는 진화 기법과 가치함수를 학습하는 기법의 차이를 보여준다. 진화 기법은 정책을 평가할 때 정책을 고정한 채 상대와 여러 번 게임을 하거나 상대의 모델을 사용하여 게임을 여러 번 시뮬레이트한다. 승리 빈도는 그 정책으로 승리할 가능성을 공평하게 예측하고, 다음에 어떤 정책을 선택할지 도움을 준. 그러나 여러 번 게임을 한 후에만 정책을 변경하며 오직 게임의 최종 결과만 고려한다. 게임 과정에서 일어난 일은 무시한다. 예를 들어, 선수가 이기면 움직임 하나하나가 우승에 얼마나 결정적이었는지 고려하지 않고 게임 중 모든 행동이 공을 인정받는다. 심지어 실제로 하지 않은 움직임에도 공이 돌아간다! 반대로 가치함수 기법은 개별 상태를 각각 평가한다. 결국 진화 기법과 가치함수 기법은 모두 정책 집합을 탐색하지만, 가치함수 학습은 경기 과정의 정보를 활용한다.

이 간단한 사례는 강화학습 기법의 몇 가지 핵심 특징을 보여준다. 먼저, 환경과 상호작용하는 학습을 강조한다. 이 경우에서 환경은 상대 선수이다. 두 번째는, 명확한 목표가 있고, 올바로 행동하려면 시간이 흐른 뒤 선택의 효과를 고려한 계획이나 통찰이 필요하다. 예를 들어, 여러 번 움직임으로 근시안적인 상대를 함정으로 몰도록 간단한 강화학습 선수를 학습할 수 있다. 강화학습 해결책은 놀랍게도 상대의 모델을 사용하거나 미래의 모든 상태와 행동을 명시적으로 탐색하지 않고도 계획과 선견지명 효과를 낸다.

이 사례가 몇 가지 강화학습의 핵심 특징을 보여주지만, 사례가 너무 단순해서 강화학습이 제한적이라는 인상을 줄 수 있다. 틱택토는 두 명이 하는 게임이지만, 강화학습은 외부의 적이 없는 경우, 즉 “자연을 상대로 한 게임”에도 적용할 수

있다. 또한, 강화학습은 턱택토처럼 행동이 에피소드로 구분되고 에피소드가 끝날 때만 보상을 받는 문제에 국한되지 않는다. 행동이 무한히 계속되거나 아무 때나 다양한 보상을 받는 경우도 적용할 수 있다. 턱택토같이 불연속 시간 스텝으로 쪼갤 수 없는 문제도 가능하다. 이론이 더 복잡하기 때문에 이 입문서에서는 다루지 않지만, 연속 시간 문제에도 동일한 원리를 적용할 수 있다.

턱택토는 상태 집합이 상대적으로 작고 유한하지만, 강화학습은 상태가 매우 많거나 심지어 무한대인 경우에도 사용할 수 있다. 예를 들어, 제리 테사우로 (Tesauro, 1992 and 1995)는 앞에서 말한 알고리즘을 인공신경망과 결합하여 상태가 10^{20} 개 정도인 백거먼(backgammon)을 학습했다. 이렇게 상태가 많으면 일부만 경험하기도 불가능하다. 테사우로가 만든 프로그램은 이전 프로그램보다 게임을 훨씬 더 잘 했고, 이제 세계 최고 인간 선수 수준으로 경기한다(16장 참고). 신경망을 사용한 프로그램은 경험을 일반화할 수 있다. 그래서 새로운 상태에서 신경망은 과거 비슷한 상태의 경험을 토대로 움직임을 선택한다. 강화학습 시스템이 이렇게 상태가 많은 문제를 잘 해결할지는 과거 경험을 적절하게 일반화하는 능력에 달렸다. 그래서 강화학습을 더한 지도학습 기법이 매우 절실하다. 신경망과 딥러닝이 (9.6절) 그 방법 중 하나다.

턱택토 예제에서 게임 규칙 외에 아무런 사전 지식 없이 학습을 시작했지만, 강화학습이라고 백지상태에서 출발해서 학습하고 지능을 얻는 것은 아니다. 반대로 다양한 방식으로 강화학습에 사전 지식을 불어넣을 수 있고, 효율적으로 학습할 때 중요할 수 있다. 또한, 여기서는 턱택토 상태를 정확히 알 수 있지만, 강화학습은 일부 상태가 숨겨졌거나 학습자가 보기기에 여러 상태가 같아 보이는 경우에도 적용할 수 있다. 그러나 이런 경우는 상당히 어렵고, 이 책에서는 중요하게 다루지 않는다.

마지막으로 턱택토 선수는 앞을 내다보고 앞으로 어떤 상태가 나올 수 있는지 알았다. 이를 위해 환경이 해보지 않은 움직임에 어떻게 반응하여 변경할지 “생각하려면” 게임 모델이 필요하다. 모델이 있는 문제가 많지만, 행동의 영향에 대한 단기 모델도 없는 경우도 있다. 강화학습은 두 경우 모두 적용할 수 있다. 모델이 필요 없지만, 모델이 있거나 모델을 학습할 수 있다면 모델을 쉽게 이용할 수 있다.

한편 환경 모델이 전혀 필요 없는 강화학습 기법도 있다. 비모델 시스템은 개별 행동에 대해 환경이 어떻게 반응할지 전혀 생각할 수 없다. 상대 측면에서 턱택토 선수는 비모델이다. 상대에 대한 모델이 없기 때문이다. 모델이 유용하려면 어느 정도 정확해야 하기 때문에 충분히 정확한 환경 모델을 만들기 어려운 문제에 막힐 때 복잡한 기법보다 비모델 기법이 유리할 수 있다. 비모델 기법은 모델기반 기법의 중요한 구성요소이기도 하다. 이 책은 여러 장에 걸쳐 비모델 기법을 다룬 다음 더 복잡한 모델기반 기법에서 어떻게 비모델 기법을 사용하는지 볼 것이다.

시스템에서 강화학습을 고수준과 저수준 모두에서 사용할 수 있다. 틱택토 선수는 게임의 기본적인 움직임만 학습하지만, 더 높은 수준에서 각각 “행동” 자체가 정교한 문제 풀이 기법인 강화학습도 가능하다. 계층적 학습 시스템은 여러 수준에서 동시에 강화학습을 할 수 있다.

연습 문제 1.1: 자가 경기(*self-play*) 앞에서 설명한 강화학습 알고리즘이 임의의 상대 대신 자기 자신과 경기를 하며 양쪽 모두 학습하는 경우를 가정하자. 어떤 일이 벌어질 것으로 생각하는가? 서로 다르게 움직이는 정책을 학습할까?

□

연습 문제 1.2: 대칭 여러 틱택토 칸은 서로 달라 보이지만 대칭 때문에 사실 같다. 이 사실을 활용하여 앞에서 설명한 학습 과정을 어떻게 보완할 수 있을까? 어떤 면에서 학습 과정이 향상되는가? 이제 다시 상대가 대칭을 활용하지 않는 경우를 가정하자. 이 경우 우리는 활용해야 하는가? 만약 활용해야 한다면, 대칭으로 동등한 자리는 반드시 가치가 같아야 하는가?

□

연습 문제 1.3: 탐욕적인 경기 강화학습 선수가 탐욕적이라고, 즉 항상 가장 가치가 높은 칸을 둔다고 가정하자. 탐욕적이지 않은 선수보다 경기를 잘 하는가 아니면 못하는가? 어떤 문제가 일어날 수 있는가?

□

연습 문제 1.4: 탐색을 통한 학습 탐색적 이동을 포함하여 모든 움직임마다 학습을 갱신한다고 가정하자. 시간이 지나면서 스텝 크기 파라미터가 적절하게 줄어든다면(그러나 탐색 확률은 줄어들지 않는다), 상태 가치는 어떤 확률에 수렴한다. 탐색적 움직임을 학습한 경우와 학습하지 않은 경우 두 확률은 어떤가? 계속 탐색한다면, 어느 편이 더 잘 학습하는가? 더 많이 이기는 쪽은 어디인가?

□

연습 문제 1.5: 다른 개선점 강화학습 선수를 향상할 다른 방법을 생각하자. 여기 틱택토 문제를 더 잘 푸는 방법이 있는가?

□

1.6 요약

강화학습은 계산을 통해 목표 지향 학습과 의사결정을 이해하고 자동화하는 접근 방식이다. 모범을 보이는 감독이나 환경의 완전한 모델에 의존하는 다른 계산적 접근방식과 다르게 강화학습은 환경과 직접 상호작용하는 에이전트의 학습을 강조한다. 우리는 강화학습이 장기 목표를 달성하기 위해 환경과 상호작용을 통해 학습할 때 발생하는 계산 문제를 심각하게 다루는 첫 번째 학문 분야라고 생각한다.

강화학습은 학습 에이전트와 환경의 상호작용을 상태와 행동과 보상 형식으로 정의한 공식 구조를 사용한다. 이 구조는 인공지능 문제의 핵심 특징을 단순하게

표현하려고 고안되었다. 인과 관계 개념, 불확실과 비결정론 개념, 명시적인 목표가 존재하는 점 등이 특징이다.

가치와 가치함수 개념은 이 책에서 다루는 대다수의 강화학습 기법의 핵심 특징이다. 우리는 정책 집합을 효율적으로 탐색하는데 가치함수가 중요하다는 입장이다. 강화학습 기법은 가치함수를 사용하기 때문에 전체 정책을 스칼라 평가하여 정책 집합을 직접 탐색하는 진화적 기법과 다르다.

1.7 강화학습의 초기 역사

강화학습의 역사는 두 가지 주된 갈래가 있다. 깊고 풍부한 두 갈래는 현대 강화학습으로 합쳐지기 전에 각자 독립적으로 연구되었다. 한 줄기는 동물 학습 심리학에서 시작한 시행착오를 통한 학습이다. 가장 오래된 인공지능 연구를 거쳐서 1980년대 초 강화학습의 부활로 이어졌다. 다른 줄기는 최적 제어 문제와 가치함수와 동적 프로그래밍을 사용한 최적 제어 풀이이다. 이 갈래는 보통 학습을 다루지 않았다. 두 줄기는 크게 서로 독립적이지만, 이 장의 틱택토 사례에서 사용한 것과 같은 시간차 기법을 다루는 덜 분명한 세 번째 갈래는 예외이다. 1980년대 말 세 줄기가 함께 모여 이 책에 나오는 현대 강화학습 분야가 되었다.

시행착오 학습을 강조한 갈래가 가장 익숙하고 여기서도 주로 다룰 것이다. 그러나 그 전에 최적 제어 갈래를 짧게 언급한다.

“최적 제어(optimal control)”란 용어는 시간에 따른 동적 시스템 행동을 최소화하는 제어기 설계 문제를 설명하려고 1950년대 말에 사용하기 시작했다. 1950년대 중반 리처드 벨만(Richard Bellman) 등이 19세기 해밀턴(Hamilton)과 자코비(Jacobi)의 이론을 확장하여 이 문제의 해결책 중 하나를 만들었다. 이 방법은 동적 시스템 상태와 가치함수 혹은 “최적 이익 함수” 개념을 사용하여 지금은 벨만 방정식(Bellman equation)이라고 부르는 함수식을 정의한다. 이 방정식을 가지고 최적 제어 문제를 푸는 기법을 동적 프로그래밍(Bellman, 1957a)이라고 부르게 된다. 또, 벨만(1957b)은 마르코프 결정 프로세스(Markovian decision processe, MDP)이라는 최적 제어 문제의 이산 확률 버전을 소개했고, 로널드 하워드(Howard, 1960)는 MDP를 위한 정책 반복법(policy iteration method)을 고안했다. 이것들은 현대 강화학습 이론과 알고리즘의 근간이 되는 핵심 요소이다.

동적 프로그래밍은 일반적인 확률 최적 제어 문제의 유일한 풀잇법으로 널리 알려졌다. 동적 프로그래밍은 벨만이 말한 (상태 변수 개수가 증가할 때 계산량이 지수적으로 증가하는) “차원의 저주”에 시달리지만, 다른 일반적인 기법보다 여전히 훨씬 더 효율적이고 폭넓게 적용할 수 있다. 동적 프로그래밍은 1950년대 후반부터 부분 관찰 MDP(Lovejoy, 1991), 여러 가지 응용(White, 1985, 1988 and 1993), 근사 기법(Rust, 1996), 비동기 기법(Bertsekas, 1982 and 1983) 등으로 널리 연구되었다. 여러 훌륭한 현대 동적 프로그래밍 교재가 있다(예, Bertsekas,

2005, 2012; Puterman, 1994; Ross, 1983; Whittle, 1982 and 1983). 브라이슨(Bryson, 1996)은 최적 제어 분야의 역사를 충실히 알려준다.

이 책에서 우리는 모든 최적 제어 기법은 어떤 의미에서 강화학습에도 가능하다고 본다. 우리는 강화학습 문제를 효율적으로 푸는 모든 방법을 강화학습 기법으로 정의한다. 그리고 강화학습 문제는 최적 제어 문제와 (특히 MDP로 기술할 수 있는 확률 최적 제어 문제와) 밀접하게 연관이 있다. 따라서 동적 프로그래밍 같은 최적 제어 풀잇법을 강화학습 기법으로 봐야 한다. 거의 모든 전통적인 기법은 제어하는 시스템을 완벽하게 알아야 하기 때문에 강화학습의 일부라고 말하기에 조금 부자연스럽다. 그러나 많은 동적 프로그래밍 알고리즘은 점진적이고 반복적이다. 학습 기법 같이 계속 근사하여 올바른 답에 도달한다. 앞으로 보겠지만 단지 겉모습만 비슷하지 않고 훨씬 근본적으로 유사하다. 완전한 지식과 불완전한 지식에 대한 이론과 해법은 너무 밀접하게 연관되었기 때문에 한 가지 주제의 일부로 함께 고려해야 한다고 생각한다.

이제 시행착오 학습 개념을 따르는 현대 강화학습을 이끈 다른 갈래로 돌아가자. 여기서는 중요한 점만 언급하고 14장에서 자세히 다룰 것이다. 미국의 심리학자 우드워스에 따르면, 시행착오 학습 개념은 1850년대 알렉산더 베인(Alexander Bain)의 “암중모색과 실험”에 의한 학습까지 거슬러 가고, 더 직접적으로는 1894년 영국의 동물행동학자 겸 심리학자 콘웨이 로이드 모간(Conway Lloyd Morgan)이 관찰한 동물 행동을 설명할 때 이 용어를 사용했다(Woodworth, 1938). 시행착오 학습이 학습 원리의 핵심이라고 처음으로 말한 사람은 아마도 에드워드 손다이크이다.

동일한 상황에서 동물이 즉시 만족하거나 곧 만족하게 되는 반응은, 다른 조건이 동일하다면, 상황과 더 강하게 연결된다. 그래서 상황이 반복하면, 반응이 반복할 가능성이 높아진다. 동물이 싫어하거나 곧 싫게 되는 반응은, 다른 조건이 동일하다면, 환경과 연관이 약해진다. 그래서 반복할수록 일어날 가능성이 낮아진다. 만족이나 불쾌함이 커질수록 연결이 강해지거나 약해진다. (Thorndike, 1911, p. 244)

강화하는 사건이 행동 선택 경향에 미치는 영향을 보여주기 때문에 손다이크는 이것을 “효과의 법칙(Law of Effect)”이라고 불렀다. 이후 손다이크는 (보상과 체벌의 효과 차이 같은) 동물 학습에 대한 축적된 자료를 더 잘 설명하도록 법칙을 변경했고, 다양한 형태의 법칙이 학습 이론가 사이에서 상당히 논란이 되었다(예, Gallistel, 2005; Herrnstein, 1970; Kimble, 1961, 1967; Mazur, 1994, 참고). 논란에도 불구하고 한두 가지 형태의 효과의 법칙은 대부분 행동의 근간이 되는 기본 원리로 널리 인정받았다(예, Hilgard and Bower, 1975; Dennett, 1978; Campbell, 1960; Cziko, 1995). 클라크 헐의 영향력 있는 학습 이론과 스키너의 실험 기법의 기초가 된다(예, Hull, 1943; Skinner, 1938).

손다이크가 효과의 법칙을 말한 이후 동물 학습 분야에서 “강화”란 용어를 널리 사용했고, 우리가 알기로는 파블로프(Pavlov) 조건반사 논문의 1927년 영어 번역에서 이런 의미로 처음 등장한다. 강화는 다른 자극 혹은 응답과 적절한 시간 관계를 가진 자극(강화물, reinforcer)을 받은 동물의 행동 패턴 강화이다. 일부 심리학자는 강화는 물론이고 약화 과정까지, 즉 사건의 생략이나 중단이 행동을 변화하는 경우를, 포함하도록 이 의미를 확장한다. 강화는 행동을 변화하고, 강화물이 사라져도 변화한 행동을 지속시킨다. 그래서 지속되는 변화 없이 동물의 관심을 끌거나 행동을 격려하는 자극은 강화물로 보지 않는다.

인공지능의 가능성에 대한 초기 발상 중에 컴퓨터로 시행착오 학습을 구현하는 아이디어가 나왔다. 앤런 튜링은 1948년 보고서에서 효과의 법칙의 연장선에서 “쾌락-고통 시스템(pleasure-pain system)” 설계를 다루었다.

무슨 행동을 할지 결정하지 않은 배치에 도달하면, 없는 데이터 자리에 임의로 선택한 적절한 항목을 기록하고 시험 삼아 적용한다. 고통 자극이 발생하면 모든 시험적 항목을 취소하고, 기쁨 자극이 발생하면 모두 그대로 둔다. (Turing, 1948)

시행착오 학습을 보여주려고 많은 독창적인 전기기계 장치가 만들어졌다. 가장 초기 장치는 토마스 로스(Ross, 1933)가 만든 간단한 미로를 통과하며 스위치 설정으로 길을 기억하는 기계였다. 1951년에는 이미 “기계 거북이”로 알려진 그레이 월터(Walter, 1950)가 단순한 학습이 가능한 거북이를 만들었다(Walter, 1951). 1952년 클로드 쇄넌은 시행착오를 통해 길을 찾는 테세우스(Theseus)라는 미로 찾기 쥐를 선보였다. 미로는 바닥 아래에 있는 자석과 릴레이를 가지고 성공한 방향을 기억했다(Shannon, 1951 and 1952). 도이치(Deutsch, 1954)는 어떤 면에서 모델기반 강화학습(8장)과 유사한 자신의 행동 이론(Deutsch, 1953)에 기반을 둔 미로 찾기 장치를 설명했다. 마빈 민스키는 박사 논문(Minsky, 1954)에서 강화 학습의 계산 모델과 함께 두뇌의 변경 가능한 시냅스 연결을 (15장) 흉내낸 SNARCs(Stochastic Neural-Analog Reinforcement Calculators)라는 구성요소로 만든 아날로그 장치를 설명했다. 멋진 웹사이트 cyberneticzoo.com에 이런 전기기계 학습 장치에 대한 정보가 풍부하다.

전기기계 학습 장치 제작은 시행착오 학습을 포함하여 다양한 종류의 학습을 수행하는 프로그래밍 디지털 컴퓨터로 이어졌다. 팔리와 클라크(Farley and Clark, 1954)는 시행착오를 통해 학습하는 신경망 학습 기계를 디지털 시뮬레이션했다. 그러나 그들은 곧 관심을 시행착오 학습에서 일반화와 패턴인식으로, 즉 강화학습에서 지도학습으로 옮겼다(Clark and Farley, 1955). 이렇게 학습 종류들의 관계를 혼동하기 시작했다. 많은 연구자이 사실은 지도학습을 연구하지만 자신은 강화학습을 연구한다고 믿곤 했다. 예를 들어, 신경망을 개척한 로젠블랫(Rosenblatt, 1962) 그리고 위드로와 호프(Widrow and Hoff, 1960)는 보상과

체별이란 언어를 사용하는 등 분명히 강화학습에서 영감을 받았지만, 패턴인식과 지각학습에 적합한 지도학습 시스템을 연구했다. 심지어 오늘날에도 일부 연구자와 교과서는 학습 종류들의 구분을 최소화하거나 모호하게 만든다. 예를 들어, 일부 신경망 교과서는 훈련 예제를 가지고 학습한 망을 설명할 때 “시행착오”란 단어를 사용한다. 망이 연결 가중치를 갱신하려고 오류 정보를 사용하기 때문에 혼동이 이해되지만, 어떤 행동이 올바른지 알지 않고 피드백을 평가하여 행동을 선택하는 시행착오 학습의 핵심 특징을 간과한다.

어느 정도는 이런 혼동 때문에 몇 가지 주목할만한 예외를 제외하고는 1960년대와 1970년대에 순수한 시행착오 학습 연구가 드물었다. 1960년대에 시행착오 학습의 공학 사례를 다루며 공학 논문에서 “강화”와 “강화학습”이란 용어가 처음 등장한다(예, Waltz and Fu, 1965; Mendel, 1966; Fu, 1970; Mendel and McLaren, 1970). 특히 민스키의 영향력 있는 “인공지능을 향한 단계(Steps Toward Artificial Intelligence)” 논문(Minsky, 1961)은 예측과 기대 그리고 그의 표현에 따르면 복잡한 강화학습 시스템을 위한 기본 크레디트-할당 문제 (성공 크레디트를 분배하는 방법) 등 시행착오 학습의 여러 쟁점을 다룬다. 이 책에서 다루는 모든 기법은 어떤 의미에서 이 문제를 풀려고 한다. 민스키의 논문은 지금도 읽을 가치가 있다.

이제 1960년대와 1970년대 계산적이고 이론적인 순수 시행착오 학습 연구와 거리가 있는 몇 가지 예외를 보자.

그중 하나는 존 안드레라는 뉴질랜드 연구자의 연구이다. 안드레(Andreae, 1963)는 환경과 상호작용하며 시행착오를 통해 학습하는 STeLLA 시스템을 개발했다. 이 시스템은 숨겨진 상태 문제를 다루기 위해 세상의 내부 모델을 가진다 (Andreae, 1969a). 내부 모델은 나중에 “내부 독백”이라고 불렸다. 안드레의 이후 연구(1977)는 교사로부터 학습을 더 강조하지만, 여전히 시행착오 요소가 있다. 불행하게도 그의 선구적인 연구는 잘 알려지지 않았고, 이후 강화학습 연구에 큰 영향을 주지 않았다.

도널드 미치(Donald Michie)의 연구는 더 영향력이 있었다. 1961년과 1963년 그는 틱택토 게임을 하는 방법을 학습하는 간단한 시행착오 학습 시스템 MENACE(Matchbox Educable Naughts and Crosses Engine)를 공개했다. 모든 게임 칸마다 성냥갑이 있고, 성냥갑에는 그 자리에서 움직일 수 있는 칸마다 다른 색으로 색칠한 알들이 들어있다. 현재 게임 칸에 해당하는 성냥갑에서 무작위로 알을 뽑아서 MENACE 움직임을 결정한다. 게임이 끝나면 MENACE 결정을 강화하거나 체벌하기 위해 게임에 사용한 알들을 성냥갑에 추가하거나 제거한다. 미치와 챔버스(Michie and Chambers, 1968)는 GLEE(Game Learning Expectimaxing Engine)라는 또 다른 틱택토 강화학습기와 BOXES라는 강화학습 제어기도 소개했다. 그들은 움직이는 카트로 막대의 균형을 잡는 학습에 BOXES를 적용했다. 막대가 쓰러지거나 카트가 선로 끝에 다다를 때만 실패 신호가 발생한다. 막대의

균형을 잡을 수 있는 교사가 설명한다고 가정한 지도학습 기법을 사용한 위드로와 스미스(Widrow and Smith, 1964)의 앞선 연구를 변형한 것이다. 미치와 챕버스의 막대균형잡이는 불완전한 지식 조건에서 강화학습의 가장 훌륭한 초기 사례 중 하나다. 우리 연구(Barto, Sutton, and Anderson, 1983; Sutton, 1984)를 시작으로 이후 강화학습 연구에 큰 영향을 미쳤다. 미치는 꾸준하게 시행착오와 학습이 인공지능에서 없어서는 안된다고 강조한다(Michie, 1974).

위드로와 굽타와 마이트라(Widrow, Gupta, and Maitra, 1973)는 위드로와 호프(Widrow and Hoff, 1960)의 최소 제곱 평균(LMS) 알고리즘을 변경하여 훈련 예제 대신 성공과 실패 신호를 학습할 수 있는 강화학습 규칙을 만들었다. 그들은 이 학습 형태를 “선택적 부트스트랩 적응”이라고 부르고, “교사와 학습”이 아니라 “비평가와 학습”이라고 표현했다. 이 규칙을 분석하여 블랙잭을 어떻게 배우는지 보여주었다. 지도학습에 엄청난 기여를 한 위드로가 강화학습에 관여한 사례이다. “비평가(critic)”란 단어는 위드로, 굽타, 마이트라 논문에서 유래했다. 이와 무관하게 뷰캐넌, 미첼, 스미스, 존슨(Buchanan, Mitchell, Smith, and Johnson, 1978)은 기계학습 맥락에서 비평가란 단어를 사용했지만 (또, Dietterich and Buchanan, 1984, 참고), 여기서는 성능 평가 이상의 일을 할 수 있는 전문가 시스템을 뜻한다.

학습 오토마타 연구는 현대 강화학습 연구로 이어지는 시행착오 갈래에 더 직접적인 영향을 미쳤다. k 암드 밴디트(k -armed bandit)로 알려진 비연합이고 순수하게 선택적인 학습 문제를 푸는 기법들이다(2장). k 암드 밴디트는 “한 손 강도(one-armed bandit)”, 즉, 손잡이가 k 개인 슬롯머신을 비유한 말이다. 학습 오토마타는 이런 문제에서 보상 확률을 향상하는 큰 메모리가 필요 없는 단순한 기계이다. 학습 오토마타는 러시아의 수학자이자 물리학자인 짜이틀린(Tsetlin)과 동료들의 1960년대 연구(짜이틀린 사후에 출판됨, 1973)에서 유래하였고, 그 후 공학에서 집중적으로 발전했다(Narendra and Thathachar, 1974 and 1989, 참고). 보상 신호를 기초로 행동 확률을 개선하는 방법인 확률 학습 오토마타 연구도 그중 하나다. 확률 학습 오토마타에 앞서 심리학에서는 1950년 윌리엄 에스테스가 학습의 확률적 이론(Estes, 1950)을 연구하기 시작했고, 심리학자로부트 부시와 통계학자 프레데릭 모스텔라(Bush and Mosteller, 1955) 등이 더 발전시켰다.

심리학에서 개발된 확률적 학습 이론을 경제학자들이 채택하여 경제학 분야에서 강화학습 연구로 이어졌다. 이 작업은 1973년 부시와 모스텔라 학습 이론을 고전 경제 모델들에 적용하면서 시작했다(Cross, 1973). 연구의 목표는 전통적인 이상적 경제 에이전트가 아니라 실제 사람 같이 행동하는 인공 에이전트 탐구다(Arthur, 1991). 이 접근 방식은 게임이론 영역에서 강화학습 연구로 커졌다. 경제학의 강화학습은 인공지능의 초기 연구와 무관하게 발전했지만, 강화학습과 게임이론은 현재 두 분야의 공통 관심 주제이다. 그러나 이 내용은 이 책의 범위를

넘어선다. 캐머러(Camerer, 2003)는 경제학의 강화학습 전통을 들려주며 노웨 등(Nowé et al., 2012)은 이 책에서 소개하는 접근 방식을 다중 에이전트로 확장한 관점에서 주제를 개관한다. 강화학습과 게임이론은 틱택토나 체커처럼 즐기는 게임을 플레이하는 프로그램에서 사용하는 강화학습과 매우 다른 주제다. 예를 들어, 지타(Szita, 2012)는 이런 강화학습과 게임 측면을 소개한다.

존 홀랜드(Holland, 1975)는 선택적 원칙에 기반한 일반 적응 시스템 이론을 시작한다. 그는 초기에 진화 기법과 k 암드 밴디트의 비연합 형태 시행착오를 주로 연구했다. 1976년과 1986년에 들어 연합과 가치 함수를 포함한 진정한 강화학습 시스템인 분류기 시스템을 소개했다. 홀랜드의 분류기 시스템의 핵심 구성요소는 항상 유전 알고리즘이었다. 유전 알고리즘으로 유용한 표현형을 진화시켰다. 여러 연구자가 분류기 시스템을 널리 개발하여 강화학습의 주된 연구 분야가 되었지만(Urbanowicz and Moore, 2009), 진화적 컴퓨팅의 다른 접근 방식과 마찬가지로 (우리는 그 자체로 강화학습이라고 여기지 않는) 유전자 알고리즘은 훨씬 많은 관심을 받았다(예, Fogel, Owens, and Walsh, 1966; Koza, 1992).

시행착오 갈래를 인공지능 분야의 강화학습으로 되살린 최대 수훈자는 핸리 클로프이다(Klopff, 1972, 1975 and 1982). 클로프는 학습 연구자들이 거의 지도 학습만 몰두하여 적응 행동의 본질적인 측면이 간과되었다고 생각했다. 클로프는 행동의 쾌락 측면, 환경에서 어떤 결과를 얻으려는 동기, 바라는 목적을 지향하고 원하지 않는 목적을 지양하는 환경 제어가 빠졌다고 보았다. 이들은 시행착오 학습의 핵심 개념이다. 이점이 지도학습과 강화학습을 구별하고 결국 강화학습에 집중하도록 도와주었기 때문에(Barto and Sutton, 1981a) 클로프의 발상은 저자들에게 특히 큰 영향을 주었다. 우리 동료들이 이룬 많은 초기 성과는 강화학습이 지도학습과 다름을 보이는 방향이었다(Barto, Sutton, and Brouwer, 1981; Barto and Sutton, 1981b; Barto and Anandan, 1985). 강화학습이 신경망 학습의 중요한 문제를, 특히 다중망 학습 알고리즘 생성 방법을, 처리할 수 있다는 연구도 있다(Barto, Anderson, and Sutton, 1982; Barto and Anderson, 1985; Barto and Anandan, 1985; Barto, 1985, 1986; Barto and Jordan, 1987). 강화학습과 신경망은 15장에서 더 다룰 것이다.

이제 강화학습 역사의 세 번째 갈래인 시간차 학습을 살펴보자. 시간차 학습 기법은 틱택토 승리 확률 같은 어떤 양을 예측하고 시간상 연속된 예측값의 차이로 동작하는 점이 특징이다. 이 갈래는 다른 두 갈래보다 작고 덜 뚜렷하지만, 시간차 기법이 새롭고 강화학습에 고유하다는 등의 이유로 강화학습에서 특히 중요한 역할을 했다.

시간차 학습은 어느 정도는 동물 학습 심리학, 특히 이차적 강화자극(*secondary reinforcer*) 개념에서 나왔다고 할 수 있다. 이차적 강화자극은 음식과 고통 같은 일차적 강화자극과 쌍을 이루는 자극으로, 결과적으로 강화 성질이 비슷하다. 아마도 민스키(1954)가 이 심리 이론이 인공 학습 시스템에 중요할 수 있다고

처음 발견했을 것이다. 아더 사무엘(Samuel, 1959)은 그의 유명한 체커 게임 프로그램에서 처음으로 시간차 개념을 가진 학습 기법을 제안하고 구현했다.

사무엘은 민스키의 연구나 동물 학습과 관계를 언급하지 않았다. 분명히 그는 평가 함수를 사용하여 컴퓨터가 체스를 두도록 프로그래밍할 수 있고 실행 중에 평가 함수를 변경하여 경기력을 향상할 수 있다는 클로드 샐런(1950)의 제안에서 영감을 받았다. (벨만도 샐런의 아이디어의 영향을 받았을지도 모르지만 증거는 없다.) 민스키(1961)는 자신의 “단계” 논문에서 자연과 인공 모두에서 이차 강화 이론과 연관성을 시사하며 사무엘의 연구를 크게 다루었다.

말했듯이 민스키와 사무엘의 연구 이후 십 년간 시행착오 학습에 대한 계산 연구가 거의 없었고, 시간차 학습 연구는 전무했다. 1972년 클로프는 시간차 학습의 중요한 요소와 함께 시행착오 학습을 제시했다. 클로프는 학습을 큰 시스템으로 확장할 수 있는 원리에 관심을 가지고, 전체 학습 시스템의 부분들이 서로를 강화하는 지역 강화 개념에 매료되었다. 그는 모든 요소(즉, 모든 뉴런)가 모든 입력을 흥분 입력은 보상으로 억제 입력은 체벌로 해석하는 “일반화된 강화” 개념을 만들었다. 현재 시간차 학습과 개념이 다르고, 뒤돌아 생각해보면 사무엘의 연구보다 더 나아갔다. 한편 클로프는 이 개념을 시행착오 학습과 결합하고, 동물 학습 심리학의 방대한 경험과 관련시켰다.

서튼(1978a,c,b)은 클로프의 발상을, 특히 동물 학습 이론과 연결하여 더 발전시켜서 시간상 연속한 예측의 변화로 학습 규칙을 설명했다. 서튼과 바토는 이 발상을 다듬어서 시간차 학습에 기반을 둔 고전적 조건 형성 심리학 모델을 만들었다(Sutton and Barto, 1981a; Barto and Sutton, 1982). 그 후 시간차 학습을 사용한 몇 가지 다른 유력한 고전적 조건 형성 심리학 모델도 나왔다(Klopf, 1988; Moore, Desmond, et al., 1986; Sutton and Barto, 1987 and 1990). 대부분 역사적 관련이 없지만, 이 시기에 나온 몇 가지 신경과학 모델도 시간차 학습 입장에서 잘 해석된다(Hawkins and Kandel, 1984; Byrne, Gingrich, and Baxter, 1990; Gelperin, Hopfield, and Tank, 1985; Tesauro, 1986; Friston et al., 1994).

우리의 초기 시간차 학습 연구는 동물 학습 이론과 클로프의 연구에 큰 영향을 받았다. 민스키 “단계” 논문과 사무엘의 체커 게임기와 관계는 시간이 흐른 후에야 알게 되었다. 그러나 1981년 무렵에 우리는 앞에서 언급한 모든 선행 연구를 시간차와 시행착오 갈래로 인식하게 되었다. 그 시기에 우리는 시행착오 학습에 시간차 학습을 사용한 액터-크리틱 아키텍처(*actor-critic architecture*) 기법을 만들고, 미치와 챕버스의 막대 균형 문제에 적용했다(Barto, Sutton, and Anderson, 1983). 이 기법은 서튼의 박사 논문(1984)에서 집중적으로 다루었고, 앤더슨의 박사 논문(1986)은 역전파 신경망을 사용하여 확장했다. 그 무렵 홀랜드(1986)는 시간차 아이디어를 그의 분류기 시스템에 직접 도입했다. 중요한 사건은 1988년 서튼이 시간차 학습을 제어와 구분하여 일반 예측 기법으로 취급한 점이다. 그

논문은 TD(λ) 알고리즘을 소개하고 일부 수렴 조건을 증명하기도 했다.

1981년 액터-크리틱 아키텍처 연구를 마무리할 때 우리는 시간차 학습 규칙을 가장 먼저 수록했다고 알려진 이언 위튼(Witten, 1977)의 논문을 발견했다. 그는 MDP를 푸는 적응 제어기의 일부로 우리가 표 TD(0)이라고 부르는 기법을 제안했다. 위튼의 연구는 안드레의 초기 STeLLA 실험과 다른 시행착오 학습 시스템을 계승했다. 그래서 위튼의 1977년 논문은 시간차 학습에 독특한 초기 기여이며 동시에 강화학습 연구의 주된 갈래(시행착오 학습과 최적 제어)를 모두 아우른다.

시간차와 최적 제어 갈래는 1989년 크리스 왓킨이 Q-학습을 개발하면서 완전히 결합한다. 이 연구는 세 강화학습 연구 갈래의 이전 연구를 확장하고 통합한다. 폴 워보스(Werbos, 1987)는 1977년부터 시행착오 학습과 동적 프로그래밍의 수렴을 주장하며 통합에 일조했다. 왓킨이 연구하던 시절 강화학습 연구가 엄청나게 성장했다. 특히 인공지능의 기계학습 분야가 성장했지만, 신경망과 인공지능도 폭넓게 발전했다. 1992년 제리 테사우로의 백거면 게임 프로그램 TD-Gammon의 놀라운 성공은 이 분야에 더 관심을 일으켰다.

이 책 초판 출판 이후 강화학습 알고리즘과 신경 시스템의 강화학습 사이의 관계를 주목한 신경과학의 하위분야가 번창했다. 여러 연구자가 지적했듯이 시간차 알고리즘의 행동과 뇌에서 도파민을 생성하는 뉴런의 동작이 신기할 정도로 비슷한 점이 그 이유이다(Friston et al., 1994; Barto, 1995a; Houk, Adams, and Barto, 1995; Montague, Dayan, and Sejnowski, 1996; Schultz, Dayan, and Montague, 1997). 15장은 강화학습의 이런 흥미로운 측면을 소개한다.

최근 강화학습 역사에 중요한 기여가 너무 많아서 이 짧은 글에 다 언급할 수 없다. 일부는 관련 장 마지막에 인용한다.

참고문헌

강화학습 일반은 Szepesvári (2010), Bertsekas and Tsitsiklis (1996), Kaelbling (1993a), Masashi Sugiyama et al. (2013) 책을 참고하라. Si et al. (2004), Powell (2011), Lewis and Liu (2012), Bertsekas (2012) 책은 제어 혹은 운용 과학 관점으로 접근한다. Cao (2009)는 확률 동적 시스템의 학습과 최적화에 대한 다른 접근 방식으로서 강화학습을 다룬다. 《Machine Learning》 저널 Sutton (1992), Kaelbling (1996), Singh (2002)은 강화학습 특별호다. Barto (1995b), Kaelbling, Littman, and Moore (1996), Keerthi and Ravindran (1997)의 개론도 유용하다. Wiering and van Otterlo (2012)가 편집한 책은 최근 발전을 잘 개관한다.

이 장에 나오는 필의 아침 예는 Agre (1988)에서 영감을 받았다. 틱택토 사례에서 사용한 일종의 시간차 기법은 6장을 참고하라.

3장에서 보듯이 이 책에서 다루는 강화학습 이론은 에이전트가 앞으로 누적 보상의 기댓값을 최대화하는 원리이다. 이성적 결정이 기대 효용을 최대화하는

von Neumann and Morgenstern (1944)의 고전 이론과 같은 맥락이다. 그러나 임의의 양의 기댓값을 최대화하는 것이 올바르지 않은 경우가 있다. 위험(*risk*)을 내재한 편차를 무시하기 때문이다. 위험에 민감한 최적화는 금융과 최적 제어처럼 과도한 위험이 파괴적인 분야에서 주로 발전했다. 위험은 강화학습에서도 중요하지만 이 책의 범위는 아니다. Heger (1994), Geibel (2001), Mihatsch and Neuneier (2002), Borkar (2002) 논문은 강화학습에서 위험을 고려하고 여기서 다루는 일부 알고리즘을 위험에 민감하게 변형한다. Coraluppi and Marcus (1999)는 우리의 강화학습 접근 방식의 기초가 되는 불연속 시간 유한 상태 마르코프 결정 프로세스 맥락에서 위험을 논한다. 우리는 사람들의 바람을 측정하는 효용 이론을 완전히 무시한다. 인간 경제 행동 이론으로서 강화학습을 다룬다면 효용 이론이 관련이 있다. 일부 강화학습 이론은 보상과 효용을 같은 것으로 본다(예, Russell and Norvig (2010)). 그러나 이 책에서 우리의 목표가 아니기 때문에 효용 이론과 연관성은 다른 분에게 남겨둔다.

제 I 편

Tabular Solution Methods

제 2 장

Multi-armed Bandits

제 3 장

Finite Markov Decision Processes

제 4 장

Dynamic Programming

제 5 장

몬테카를로 기법

$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

$$\pi(s) \doteq \arg \max_a q(s, a). \quad (5.1)$$

$$\begin{aligned} q_{\pi_k}(s, \pi_{k+1}(s)) &= q_{\pi_k}(s, \arg \max_a q_{\pi_k}(s, a)) \\ &= \max_a q_{\pi_k}(s, a) \\ &\geq q_{\pi_k}(s, \pi_k(s)) \\ &\geq v_{\pi_k}(s). \end{aligned}$$

$$\begin{aligned} q_\pi(s, \pi'(s)) &= \sum_a \pi'(a|s) q_\pi(s, a) \\ &= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon) \max_a q_\pi(s, a) \\ &\geq \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon) \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|\mathcal{A}(s)|}}{1 - \epsilon} q_\pi(s, a) \end{aligned} \quad (5.2)$$

(the sum is a weighted average with nonnegative weights summing to 1, and as such it must be less than or equal to the largest number averaged)

$$\begin{aligned}
&= \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) - \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a q_\pi(s, a) + \sum_a \pi(a|s) q_\pi(s, a) \\
&= v_\pi(s).
\end{aligned}$$

$$\begin{aligned}
\tilde{v}_*(s) &= (1 - \epsilon) \max_a \tilde{q}_*(s, a) + \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a \tilde{q}_*(s, a) \\
&= (1 - \epsilon) \max_a \sum_{s', r} p(s', r|s, a) \left[r + \gamma \tilde{v}_*(s') \right] \\
&\quad + \frac{\epsilon}{|\mathcal{A}(s)|} \sum_a \sum_{s', r} p(s', r|s, a) \left[r + \gamma \tilde{v}_*(s') \right].
\end{aligned}$$

제 6 장

Temporal-Difference Learning

제 7 장

Multi-step Bootstrapping

제 8 장

Planning and Learning with Tabular Methods

제 II 편

Approximate Solution Methods

제 9 장

On-policy Prediction with Approximation

- 9.1 Value-function Approximation
- 9.2 The Prediction Objective (MSVE)
- 9.3 Stochastic-gradient and Semi-gradient Methods
- 9.4 Linear Methods
- 9.5 Feature Construction for Linear Methods
- 9.6 Nonlinear Function Approximation: Artificial Neural Networks

제 10 장

On-policy Control with Approximation

제 11 장

Off-policy Methods with Approximation

제 12 장

Eligibility Traces

제 13 장

Policy Gradient Methods

제 III 편

Looking Deeper

제 14 장

Psychology

제 15 장

Neuroscience

제 16 장

Applications and Case Studies

참고 문헌

- Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. ACM.
- Abramson, B. (1990). Expected-outcome: A general model of static evaluation. *IEEE transactions on pattern analysis and machine intelligence* 12(2):182–193.
- Adams, C. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology*, 34(2):77–98.
- Adams, C. D. and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology* 33(2):109–121.
- Adams, R. A., Huys, Q. J. M., and Roiser, J. P. (2015). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *Journal of Neurology, Neurosurgery & Psychiatry*, doi:10.1136/jnnp-2015-310737.
- Agrawal, R. (1995). Sample mean based index policies with $O(\log n)$ regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27:1054–1078.
- Agre, P. E. (1988). *The Dynamic Structure of Everyday Life*. Ph.D. thesis, Massachusetts Institute of Technology. AI-TR 1085, MIT Artificial Intelligence Laboratory. (cit. on p. 23).
- Agre, P. E., Chapman, D. (1990). What are plans for? *Robotics and Autonomous Systems*, 6:17–34.
- Aizerman, M. A., Braverman, E. I., and Rozonoer, L. I. (1964). Probability problem of pattern recognition learning and potential functions method. *Avtomat. i Telemekh* 25(9):1307–1323.
- Albus, J. S. (1971). A theory of cerebellar function. *Mathematical Biosciences*, 10:25–61.
- Albus, J. S. (1981). *Brain, Behavior, and Robotics*. Byte Books, Peterborough, NH.
- Aleksandrov, V. M., Sysoev, V. I., Shemeneva, V. V. (1968). Stochastic optimization of systems. *Izv. Akad. Nauk SSSR, Tekh. Kibernetika*, 14–19.
- Amari, S. I. (1998). Natural gradient works efficiently in learning. *Neural Computation* 10(2), 251–276.
- An, P.-C. E. (1991). *An Improved Multi-dimensional CMAC Neural network: Receptive Field Function and Placement* (Doctoral dissertation, PhD Thesis, Dept. Electrical and Computer Engineering, New Hampshire Univ., New Hampshire, USA).
- An, P. C. E., Miller, W. T., Parks, P. C. (1991). Design improvements in associative memories for cerebellar model articulation controllers (CMAC). *Artificial Neural Networks*, pp. 1207–1210, Elsvier North-Holland.
- Anderson, C. W. (1986). *Learning and Problem Solving with Multilayer Connectionist Systems*. Ph.D. thesis, University of Massachusetts, Amherst. (cit. on p. 22).

- Anderson, C. W. (1987). Strategy learning with multilayer connectionist representations. *Proceedings of the Fourth International Workshop on Machine Learning*, pp. 103–114. Morgan Kaufmann, San Mateo, CA.
- Anderson, C. W. (1989). Learning to control an inverted pendulum using neural networks. *IEEE Control Systems Magazine* 9(3):31–37.
- Anderson, J. A., Silverstein, J. W., Ritz, S. A., Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84:413–451.
- Andreae, J. H. (1963). STELLA: A scheme for a learning machine. In *Proceedings of the 2nd IFAC Congress, Basle*, pp. 497–502. Butterworths, London. (cit. on p. 19).
- Andreae, J. H. (1969a). A learning machine with monologue. *International Journal of Man-Machine Studies*, 1:1–20. (cit. on p. 19).
- Andreae, J. H. (1969b). Learning machines—a unified view. In A. R. Meetham and R. A. Hudson (eds.), *Encyclopedia of Information, Linguistics, and Control*, pp. 261–270. Pergamon, Oxford.
- Andreae, J. H. (1977). *Thinking with the Teachable Machine*. Academic Press, London. (cit. on p. 19).
- Arthur, W. B. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *The American Economic Review* 81(2):353–359. (cit. on p. 20).
- Atkeson, C. G. (1992). Memory-based approaches to approximating continuous functions. In *Sante Fe Institute Studies in the Sciences of Complexity*, Proceedings Vol. 12, pp. 521–521. Addison-Wesley Publishing Co.
- Atkeson, C. G., Moore, A. W., and Schaal, S. (1997). Locally weighted learning. *Artificial Intelligence Review* 11:11–73.
- Auer, P., Cesa-Bianchi, N., Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2–3):235–256.
- Bae, J., Chhatbar, P., Francis, J. T., Sanchez, J. C., and Principe, J. C. (2011). Reinforcement learning via kernel temporal difference. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5662–5665. IEEE.
- Baird, L. C. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 30–37. Morgan Kaufmann, San Francisco.
- Baird, L. C. and Klopf, A. H. (1993). Reinforcement learning with high-dimensional, continuous actions. Wright Laboratory, Wright-Patterson Air Force Base, Tech. Rep. WL-TR-93-1147.
- Baird, L., Moore, A. W. (1999). Gradient descent for general reinforcement learning. *Advances in Neural Information Processing Systems*, pp. 968–974.
- Baldassarre, G. and Mirolli M., editors (2013). *Intrinsically Motivated Learning in Natural and Artificial Systems*. Springer-Verlag, Berlin.
- Balke, A., Pearl, J. (1994). Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence* (pp. 46–54). Morgan Kaufmann.
- Bao, G., Cassandras, C. G., Djaferis, T. E., Gandhi, A. D., Looze, D. P. (1994). Elevator dispatchers for down peak traffic. Technical report. ECE Department, University of Massachusetts, Amherst.

- Baras, D. and Meir, R. (2007). Reinforcement learning, spike-time-dependent plasticity, and the BCM rule. *Neural Computation*, 19(8):2245–2279.
- Barnard, E. (1993). Temporal-difference methods and Markov models. *IEEE Transactions on Systems, Man, and Cybernetics* 23:357–365.
- Barnhill, R. E. (1977). Representation and approximation of surfaces. *Mathematical Software* 3:69–120.
- Barreto, A. S., Precup, D., and Pineau, J. (2011). Reinforcement learning using kernel-based stochastic factorization. In *Advances in Neural Information Processing Systems*, pp. 720–728.
- Bartlett, P. L. and Baxter, J. (1999). Hebbian synaptic modifications in spiking neurons that learn. Technical report, Research School of Information Sciences and Engineering, Australian National University.
- Bartlett, P. L. and Baxter, J. (2000). A biologically plausible and locally optimal learning algorithm for spiking neurons. Rapport technique, Australian National University.
- Barto, A. G. (1985). Learning by statistical cooperation of self-interested neuron-like computing elements. *Human Neurobiology*, 4:229–256. (cit. on p. 21).
- Barto, A. G. (1986). Game-theoretic cooperativity in networks of self-interested units. In J. S. Denker (ed.), *Neural Networks for Computing*, pp. 41–46. American Institute of Physics, New York. (cit. on p. 21).
- Barto, A. G. (1989). From chemotaxis to cooperativity: Abstract exercises in neuronal learning strategies. In Durbin, R., Maill, R., and Mitchison, G., editors, *The Computing Neuron*, pages 73–98. Addison-Wesley, Reading, MA.
- Barto, A. G. (1990). Connectionist learning for control: An overview. In T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 5–58. MIT Press, Cambridge, MA.
- Barto, A. G. (1991). Some learning tasks from a control perspective. In L. Nadel and D. L. Stein (eds.), *1990 Lectures in Complex Systems*, pp. 195–223. Addison-Wesley, Redwood City, CA.
- Barto, A. G. (1992). Reinforcement learning and adaptive critic methods. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 469–491. Van Nostrand Reinhold, New York.
- Barto, A. G. (1995a). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), *Models of Information Processing in the Basal Ganglia*, pp. 215–232. MIT Press, Cambridge, MA. (cit. on p. 23).
- Barto, A. G. (1995b). Reinforcement learning. In M. A. Arbib (ed.), *Handbook of Brain Theory and Neural Networks*, pp. 804–809. MIT Press, Cambridge, MA. (cit. on p. 23).
- Barto, A. G. (2011). Adaptive real-time dynamic programming. In Sammut, C. and Webb, G. I. (Eds.) *Encyclopedia of machine learning*, pp. 19–22. Springer Science and Business Media.
- Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems*, pp. 17–47. Springer Berlin Heidelberg.
- Barto, A. G., Anandan, P. (1985). Pattern recognizing stochastic learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:360–375. (cit. on p. 21).
- Barto, A. G., Anderson, C. W. (1985). Structural learning in connectionist systems. In *Program of the Seventh Annual Conference of the Cognitive Science Society*, pp. 43–54. (cit. on p. 21).

- Barto, A. G., Anderson, C. W., Sutton, R. S. (1982). Synthesis of nonlinear control surfaces by a layered associative search network. *Biological Cybernetics*, 43:175–185. (cit. on p. 21).
- Barto, A. G., Bradtke, S. J., Singh, S. P. (1991). Real-time learning and control using asynchronous dynamic programming. Technical Report 91-57. Department of Computer and Information Science, University of Massachusetts, Amherst.
- Barto, A. G., Bradtke, S. J., Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72:81–138.
- Barto, A. G., Duff, M. (1994). Monte Carlo matrix inversion and reinforcement learning. In J. D. Cohen, G. Tesauro, and J. Alspector (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference*, pp. 687–694. Morgan Kaufmann, San Francisco.
- Barto, A. G., Jordan, M. I. (1987). Gradient following without back-propagation in layered networks. In M. Caudill and C. Butler (eds.), *Proceedings of the IEEE First Annual Conference on Neural Networks*, pp. II629–II636. SOS Printing, San Diego, CA. (cit. on p. 21).
- Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems* 13(4):341–379.
- Barto, A. G., Singh, S., and Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *International Conference on Developmental Learning (ICDL)*, LaJolla, CA.
- Barto, A. G., Sutton, R. S. (1981a). Goal seeking components for adaptive intelligence: An initial assessment. Technical Report AFWAL-TR-81-1070. Air Force Wright Aeronautical Laboratories/Avionics Laboratory, Wright-Patterson AFB, OH. (cit. on p. 21).
- Barto, A. G., Sutton, R. S. (1981b). Landmark learning: An illustration of associative search. *Biological Cybernetics*, 42:1–8. (cit. on p. 21).
- Barto, A. G., Sutton, R. S. (1982). Simulation of anticipatory responses in classical conditioning by a neuron-like adaptive element. *Behavioural Brain Research*, 4:221–235. (cit. on p. 22).
- Barto, A. G., Sutton, R. S., Anderson, C. W. (1983). Neuronlike elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:835–846. Reprinted in J. A. Anderson and E. Rosenfeld (eds.), *Neurocomputing: Foundations of Research*, pp. 535–549. MIT Press, Cambridge, MA, 1988. (cit. on pp. 20, 22).
- Barto, A. G., Sutton, R. S., Brouwer, P. S. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, 40:201–211. (cit. on p. 21).
- Barto, A. G., Sutton, R. S., and Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel and J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 539–602. MIT Press, Cambridge, MA.
- Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2012). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.
- Bellemare, M. G., Veness, J., and Bowling, M. (2012). Investigating contingency awareness using Atari 2600 games. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2012)*, pages 864–871, Palo Alto, CA. The AAAI Press.

- Bellman, R. E. (1956). A problem in the sequential design of experiments. *Sankhya*, 16:221–229.
- Bellman, R. E. (1957a). *Dynamic Programming*. Princeton University Press, Princeton. (cit. on p. 16).
- Bellman, R. E. (1957b). A Markov decision process. *Journal of Mathematical Mechanics*, 6:679–684. (cit. on p. 16).
- Bellman, R. E., Dreyfus, S. E. (1959). Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, 13:247–251.
- Bellman, R. E., Kalaba, R., Kotkin, B. (1973). Polynomial approximation—A new computational technique in dynamic programming: Allocation processes. *Mathematical Computation*, 17:155–161.
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–27.
- Bengio, Y., Courville, A. C., and Vincent, P. (2012). Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR* 1, arXiv 1206.5538.
- Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM* 18(9):509–517.
- Berg, H. C. (1975). Chemotaxis in bacteria. *Annual review of biophysics and bioengineering*, 4(1):119–136.
- Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36. English translation of the 1738 paper.
- Berns, G. S., McClure, S. M., Pagnoni, G., and Montague, P. R. (2001). Predictability modulates human brain response to reward. *The journal of neuroscience*, 21(8):2793–2798.
- Berridge, K. C. and Kringlebach, M. L. (2008). Affective neuroscience of pleasure: reward in humans and animals. *Psychopharmacology*, 199(3):457–480.
- Berridge, K. C. and Robinson, T. E. (1998). What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369.
- Berry, D. A., Fristedt, B. (1985). *Bandit Problems*. Chapman and Hall, London.
- Bertsekas, D. P. (1982). Distributed dynamic programming. *IEEE Transactions on Automatic Control*, 27:610–616. (cit. on p. 16).
- Bertsekas, D. P. (1983). Distributed asynchronous computation of fixed points. *Mathematical Programming*, 27:107–120. (cit. on p. 16).
- Bertsekas, D. P. (1987). *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, Englewood Cliffs, NJ.
- Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control, Volume 1*, third edition. Athena Scientific, Belmont, MA. (cit. on p. 16).
- Bertsekas, D. P. (2012). *Dynamic Programming and Optimal Control, Volume 2: Approximate Dynamic Programming*, fourth edition. Athena Scientific, Belmont, MA. (cit. on pp. 17, 23).
- Bertsekas, D. P. (2013). Rollout algorithms for discrete optimization: A survey. In *Handbook of Combinatorial Optimization*, pp. 2989–3013. Springer New York.
- Bertsekas, D. P., Tsitsiklis, J. N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, Englewood Cliffs, NJ.
- Bertsekas, D. P., Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA. (cit. on p. 23).

- Bertsekas, D. P., Tsitsiklis, J. N., and Wu, C. (1997). Rollout algorithms for combinatorial optimization. *Journal of Heuristics* 3(3):245–262.
- Bertsekas, D. P., Yu, H. (2009). Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics*, 227(1):27–50.
- Bhat, N., Farias, V., and Moallemi, C. C. (2012). Non-parametric approximate dynamic programming via the kernel method. In *Advances in Neural Information Processing Systems*, pp. 386–394.
- Bhatnagar, S., Sutton, R., Ghavamzadeh, M., Lee, M. (2009). Natural actor-critic algorithms. *Automatica* 45(11).
- Biermann, A. W., Fairfield, J. R. C., Beres, T. R. (1982). Signature table systems and learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 12:635–648.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon, Oxford.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, 4:113–134.
- Boakes, R. A. and Costa, D. S. J. (2014). Temporal contiguity in associative learning: Interference and decay from an historical perspective. *Journal of Experimental Psychology: Animal Learning and Cognition*, 40(4):381–400.
- Booker, L. B. (1982). *Intelligent Behavior as an Adaptation to the Task Environment*. Ph.D. thesis, University of Michigan, Ann Arbor.
- Boone, G. (1997). Minimum-time control of the acrobot. In *1997 International Conference on Robotics and Automation*, pp. 3281–3287. IEEE Robotics and Automation Society.
- Borkar, V. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2), 294–311. (cit. on p. 24).
- Bottou, L. and Vapnik, V. (1992). Local learning algorithms. *Neural Computation* 4(6):888–900.
- Boutilier, C., Dearden, R., Goldszmidt, M. (1995). Exploiting structure in policy construction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1104–1111. Morgan Kaufmann.
- Boyan, J. A. (1999). Least-squares temporal difference learning. *International Conference on Machine Learning* 16, pp. 49–56.
- Boyan, J. A. (2002). Technical update: Least-squares temporal difference learning. *Machine Learning* 49:233–246.
- Boyan, J. A., Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. In G. Tesauro, D. S. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 369–376. MIT Press, Cambridge, MA.
- Bradtko, S. J. (1993). Reinforcement learning applied to linear quadratic regulation. In S. J. Hanson, J. D. Cowan, and C. L. Giles (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1992 Conference*, pp. 295–302. Morgan Kaufmann, San Mateo, CA.
- Bradtko, S. J. (1994). *Incremental Dynamic Programming for On-Line Adaptive Optimal Control*. Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 94-62.
- Bradtko, S. J., Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57.

- Bradtko, S. J., Duff, M. O. (1995). Reinforcement learning methods for continuous-time Markov decision problems. In G. Tesauro, D. Touretzky, and T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 393–400. MIT Press, Cambridge, MA.
- Bradtko, S. J., Ydstie, B. E., Barto, A. G. (1994). Adaptive linear quadratic control using policy iteration. In *Proceedings of the American Control Conference*, pp. 3475–3479. American Automatic Control Council, Evanston, IL.
- Brafman, R. I., Tennenholtz, M. (2003). R-max – a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3, 213–231.
- Breiter, H. C., Aharon, I., Kahneman, D., Dale, A., and Shizgal, P. (2001). Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron*, 30(2):619–639.
- Breland, K. and Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, 16(11):681–684.
- Bridle, J. S. (1990). Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimates of parameters. In D. S. Touretzky (ed.), *Advances in Neural Information Processing Systems: Proceedings of the 1989 Conference*, pp. 211–217. Morgan Kaufmann, San Mateo, CA.
- Bromberg-Martin, E. S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, 104(2):1068–1076.
- Broomhead, D. S., Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355.
- Brown, J., Bullock, D., and Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *The Journal of Neuroscience*, 19(23):10502–10511.
- Browne, C.B., Powley, E., Whitehouse, D., Lucas, S.M., Cowling, P.I., Rohlfschagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games* 4(1):1–43.
- Bryson A. E., Jr. (1996). Optimal control—1950 to 1985. *IEEE Control Systems*, 13(3):26–33. (cit. on p. 17).
- Buchanan, B. G., Mitchell, T., Smith, R. G., and Johnson Jr., C. R. (1978). Models of learning systems. *Encyclopediad of Computer Science and technology*, 11. (cit. on p. 20).
- Buhusi, C. V. and Schmajuk, N. A. (1999). Timing in simple conditioning and occasion setting: A neural network approach. *Behavioural processes*, 45(1):33–57.
- Burke, C. J., Dreher, J.-C., Seymour, B., and Tobler, P. N. (2014). State-dependent value representation: evidence from the striatum. *Frontiers in Neuroscience*, 8.
- Bush, R. R., Mosteller, F. (1955). *Stochastic Models for Learning*. Wiley, New York. (cit. on p. 20).
- Buşoniu, L., Lazaric, A., Ghavamzadeh, M., Munos, R., Babuška, R., and De Schutter, B. (2012). Least-squares methods for policy iteration. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 75–109. Springer Berlin Heidelberg.
- Byrne, J. H., Gingrich, K. J., Baxter, D. A. (1990). Computational capabilities of single neurons: Relationship to simple forms of associative and nonassociative learning in

- aplysia*. In R. D. Hawkins and G. H. Bower (eds.), *Computational Models of Learning*, pp. 31–63. Academic Press, New York. (cit. on p. 22).
- Calabresi, P., Picconi, B., Tozzi, A., and Filippo, M. D. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neuroscience*, 30(5):211–219.
- Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press. (cit. on p. 21).
- Campbell, D. T. (1960). Blind variation and selective survival as a general strategy in knowledge-processes. In M. C. Yovits and S. Cameron (eds.), *Self-Organizing Systems*, pp. 205–231. Pergamon, New York. (cit. on p. 17).
- Cao, X. R. (2009). Stochastic learning and optimization—A sensitivity-based approach. *Annual Reviews in Control* 33(1):11–24. (cit. on p. 23).
- Carlström, J., Nordström, E. (1997). Control of self-similar ATM call traffic by reinforcement learning. In *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications 3*, pp. 54–62. Erlbaum, Hillsdale, NJ.
- Chapman, D., Kaelbling, L. P. (1991). Input generalization in delayed reinforcement learning: An algorithm and performance comparisons. In *Proceedings of the Twelfth International Conference on Artificial Intelligence*, pp. 726–731. Morgan Kaufmann, San Mateo, CA.
- Chow, C.-S., Tsitsiklis, J. N. (1991). An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control*, 36:898–914.
- Chrisman, L. (1992). Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 183–188. AAAI/MIT Press, Menlo Park, CA.
- Christensen, J., Korf, R. E. (1986). A unified theory of heuristic evaluation functions and its application to learning. In *Proceedings of the Fifth National Conference on Artificial Intelligence*, pp. 148–152. Morgan Kaufmann, San Mateo, CA.
- Cichosz, P. (1995). Truncating temporal differences: On the efficient implementation of TD(λ) for reinforcement learning. *Journal of Artificial Intelligence Research*, 2:287–318.
- Claridge-Chang, A., Roorda, R. D., Vrontou, E., Sjulson, L., Li, H., Hirsh, J., and Miesenböck, G. (2009). Writing memories with light-addressable reinforcement circuitry. *Cell*, 139(2):405–415.
- Clark, R. E. and Squire, L. R. (1998). Classical conditioning and brain systems: the role of awareness. *Science*, 280(5360):77–81.
- Clark, W. A., Farley, B. G. (1955). Generalization of pattern recognition in a self-organizing system. In *Proceedings of the 1955 Western Joint Computer Conference*, pp. 86–91. (cit. on p. 18).
- Clouse, J. (1996). *On Integrating Apprentice Learning and Reinforcement Learning* TITLE2. Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 96-026.
- Clouse, J., Utgoff, P. (1992). A teaching method for reinforcement learning systems. In *Proceedings of the Ninth International Machine Learning Conference*, pp. 92–101. Morgan Kaufmann, San Mateo, CA.
- Cobo, L. C., Zang, P., Isbell, C. L., and Thomaz, A. L. (2011). Automatic state abstraction from demonstration. In *IJCAI Proceedings: International Joint Conference on Artificial Intelligence*, volume 22, page 1243.

- Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., and Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* 482(7383):85–88.
- Colombetti, M., Dorigo, M. (1994). Training agent to perform sequential behavior. *Adaptive Behavior*, 2(3):247–275.
- Connell, J. (1989). A colony architecture for an artificial creature. Technical Report AI-TR-1151. MIT Artificial Intelligence Laboratory, Cambridge, MA.
- Connell, J., Mahadevan, S. (1993). *Robot Learning*. Kluwer Academic, Boston.
- Connell, M. E. and Utgoff, P. E. (1987). Learning to control a dynamic physical system. *Computational intelligence* 3(1):330–337.
- Contreras-Vidal, J. L. and Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of computational neuroscience*, 6(3):191–214.
- Coraluppi, S. P. and Marcus, S. I. (1999). Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35:301–309 (cit. on p. 24).
- Coulom, R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. In *Proceedings of the 5th International Conference on Computers and Games*, pp. 72–83.
- Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Science*, 10(7):294–300.
- Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press, Cambridge.
- Crites, R. H. (1996). *Large-Scale Dynamic Optimization Using Teams of Reinforcement Learning Agents*. Ph.D. thesis, University of Massachusetts, Amherst.
- Crites, R. H., Barto, A. G. (1996). Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1017–1023. MIT Press, Cambridge, MA.
- Cross, J. G. (1973). A stochastic learning model of economic behavior. *The Quarterly Journal of Economics* 87(2):239–266. (cit. on p. 20).
- Crow, T. J. (1968). Cortical synapses and reinforcement: a hypothesis. *Nature*, 219:736–737.
- Curtiss, J. H. (1954). A theoretical comparison of the efficiencies of two classical methods and a Monte Carlo method for computing one component of the solution of a set of linear algebraic equations. In H. A. Meyer (ed.), *Symposium on Monte Carlo Methods*, pp. 191–233. Wiley, New York.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.
- Cziko, G. (1995). *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution*. MIT Press, Cambridge, MA. (cit. on p. 17).
- Daniel, J. W. (1976). Splines and efficiency in dynamic programming. *Journal of Mathematical Analysis and Applications*, 54:402–407.
- Dann, C., Neumann, G., Peters, J. (2014). Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research* 15:809–883.
- Daw, N. D., Courville, A. C., and Touretzky, D. S. (2003). Timing and partial observability in the dopamine system. In *Advances in neural information processing systems*, pages 99–106.

- Daw, N. D., Courville, A. C., and Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, 18(7):1637–1677.
- Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.
- Daw, N. D. and Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition*, 26(5):593–620.
- Dayan, P. (1991). Reinforcement comparison. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, and G. E. Hinton (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, pp. 45–51. Morgan Kaufmann, San Mateo, CA.
- Dayan, P. (1992). The convergence of TD(λ) for general λ . *Machine Learning*, 8:341–362.
- Dayan, P. (2008). The role of value systems in decision making. In Engel, C. and Singer, W., editors, *Better Than Conscious?: Decision Making, the Human Mind, and Implications For Institutions (Strüngmann Forum Reports)*, pages 51–70. MIT Press, Cambridge, MA.
- Dayan, P. and Abbott, L. F. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA.
- Dayan, P. and Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revaluation. *Cognitive, Affective, & Behavioral Neuroscience*, 14(2):473–492.
- Dayan, P. and Hinton, G. E. (1993). Feudal reinforcement learning. In S. J. Hanson, J. D. Cohen, and C. L. Giles (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1992 Conference*, pp. 271–278. Morgan Kaufmann, San Mateo, CA.
- Dayan, P. and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196.
- Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks* 19(8):1153–1160.
- Dayan, P. and Sejnowski, T. (1994). TD(λ) converges with probability 1. *Machine Learning*, 14:295–301.
- De Asis, K., Hernandez-Garcia, J. F., Holland, G. Z., and Sutton, R. S. (2017). Multi-step Reinforcement Learning: A Unifying Algorithm. arXiv preprint arXiv:1703.01327.
- Dean, T., Lin, S.-H. (1995). Decomposition techniques for planning in stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1121–1127. Morgan Kaufmann. See also Technical Report CS-95-10, Brown University, Department of Computer Science, 1995.
- Degrif, T., White, M., Sutton, R. S. (2012). Off-policy actor-critic. *Proceedings of the 29th International Conference on Machine Learning*.
- DeJong, G., Spong, M. W. (1994). Swinging up the acrobot: An example of intelligent control. In *Proceedings of the American Control Conference*, pp. 2158–2162. American Automatic Control Council, Evanston, IL.
- Denardo, E. V. (1967). Contraction mappings in the theory underlying dynamic programming. *SIAM Review*, 9:165–177.
- Dennett, D. C. (1978). *Brainstorms*, pp. 71–89. Bradford/MIT Press, Cambridge, MA. (cit. on p. 17).
- Derthick, M. (1984). Variations on the Boltzmann machine learning algorithm. Carnegie-Mellon University Department of Computer Science Technical Report No. CMU-CS-84-120.

- Deutsch, J. A. (1953). A new type of behaviour theory. *British Journal of Psychology. General Section*, 44(4):304–317. (cit. on p. 18).
- Deutsch, J. A. (1954). A machine with insight. *Quarterly Journal of Experimental Psychology*, 6(1):6–11. (cit. on p. 18).
- Dick, T. (2015). *Policy Gradient Reinforcement Learning Without Regret*. MSc Thesis, University of Alberta.
- Dickinson, A. (1980). *Contemporary Animal Learning Theory*. Cambridge University Press, Cambridge.
- Dickinson, A. (1985). Actions and habits: the development of behavioral autonomy. *Phil. Trans. R. Soc. Lond. B*, 308(1135):67–78.
- Dickinson, A. and Balleine, B. W. (2002). The role of learning in motivation. In Gallistel, C. R., editor, *Stevens handbook of experimental psychology*, volume 3, pages 497–533. Wiley, NY.
- Dietterich, T. G. and Buchanan, B. G. (1984). The role of the critic in learning systems. In Selfridge, O. G., Rissland, E. L., and Arbib, M. A., editors, *Adaptive Control of Ill-Defined Systems*, pages 127–147. Plenum Press, NY. Proceedings of the NATO Advanced Research Institute on Adaptive Control of Ill-defined Systems, NATO Conference Series II, Systems Science, Vol. 16. (cit. on p. 20).
- Dietterich, T. G., Flann, N. S. (1995). Explanation-based learning and reinforcement learning: A unified view. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 176–184. Morgan Kaufmann, San Francisco.
- Dietterich, T. G. and Wang, X. (2002). Batch value function approximation via support vectors. In *Advances in Neural Information Processing Systems 14*, pp. 1491–1498. Cambridge, MA: MIT Press.
- Diuk, C., Cohen, A., and Littman, M. L. (2008). An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on machine learning*, pages 240–247. ACM New York, NY.
- Dolan, R. J. and Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2):312–325.
- Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22:1–7.
- Donahoe, J. W. and Burgos, J. E. (2000). Behavior analysis and revaluation. *Journal of the Experimental Analysis of Behavior*, 74(3):331–346.
- Dorigo, M. and Colombetti, M. (1994). Robot shaping: Developing autonomous agents through learning. *Artificial Intelligence*, 71(2):321–370.
- Doya, K. (1996). Temporal difference learning in continuous time and space. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1073–1079. MIT Press, Cambridge, MA.
- Doya, K. and Sejnowski, T. J. (1995). A novel reinforcement model of birdsong vocalization learning. In Tesauro, G., Touretzky, D. S., and Leen, T., editors, *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pages 101–108, Cambridge, MA. MIT Press.
- Doya, K. and Sejnowski, T. J. (1998). A computational model of birdsong learning by auditory experience and auditory feedback. In *Central auditory processing and neural modeling*, pages 77–88. Springer US.

- Doyle, P. G., Snell, J. L. (1984). *Random Walks and Electric Networks*. The Mathematical Association of America. Carus Mathematical Monograph 22.
- Dreyfus, S. E., Law, A. M. (1977). *The Art and Theory of Dynamic Programming*. Academic Press, New York.
- Duda, R. O., Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley, New York.
- Duff, M. O. (1995). Q-learning for bandit problems. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 209–217. Morgan Kaufmann, San Francisco.
- Egger, D. M. and Miller, N. E. (1962). Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology*, 64:97–104.
- Eshel, N., Bukwich, M., Rao, V., Hemmeler, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* 525(7568):243–246.
- Eshel, N., Tian, J., Bukwich, M., and Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience* 19(3):479–486.
- Estes, W. K. (1943). Discriminative conditioning. I. A discriminative property of conditioned anticipation. *Journal of Experimental Psychology* 32(2):150–155.
- Estes, W. K. (1948). Discriminative conditioning. II. Effects of a Pavlovian conditioned stimulus upon a subsequently established operant response. *Journal of experimental psychology* 38(2):173–177.
- Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, 57:94–107. (cit. on p. 20).
- Farley, B. G., Clark, W. A. (1954). Simulation of self-organizing systems by digital computer. *IRE Transactions on Information Theory*, 4:76–84. (cit. on p. 18).
- Farries, M. A. and Fairhall, A. L. (2007). Reinforcement learning with modulated spike timing-dependent synaptic plasticity. *Journal of neurophysiology*, 98(6):3648–3665.
- Feldbaum, A. A. (1965). *Optimal Control Systems*. Academic Press, New York.
- Finch, G. and Culler, E. (1934). Higher order conditioning with constant motivation. *The American Journal of Psychology*, 596–602.
- Finnsson, H., Björnsson, Y. (2008). Simulation-based approach to general game playing. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 259–264.
- Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898–1902.
- Fiorillo, C. D., Yun, S. R., and Song, M. R. (2013). Diversity and homogeneity in responses of midbrain dopamine neurons. *The Journal of Neuroscience*, 33(11):4693–4709.
- Florian, R. V. (2007). Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19(6):1468–1502.
- Fogel, L. J., Owens, A. J., Walsh, M. J. (1966). *Artificial intelligence through simulated evolution*. John Wiley and Sons. (cit. on p. 21).
- Frey, U. and Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(6616):533–536.
- Friedman, J. H., Bentley, J. L., and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* 3(3):209–226.

- Friston, K. J., Tononi, G., Reeke, G. N., Sporns, O., Edelman, G. M. (1994). Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, 59:229–243. (cit. on pp. 22, 23).
- Fu, K. S. (1970). Learning control systems—Review and outlook. *IEEE Transactions on Automatic Control*, 15:210–221. (cit. on p. 19).
- Galanter, E., Gerstenhaber, M. (1956). On thought: The extrinsic theory. *Psychological Review*, 63:218–227.
- Gallant, S. I. (1993). *Neural Network Learning and Expert Systems*. MIT Press, Cambridge, MA.
- Gallistel, C. R. (2005). Deconstructing the law of effect. *Games and Economic Behavior* 52(2), 410–423. (cit. on p. 17).
- Gällmo, O., Asplund, L. (1995). Reinforcement learning by construction of hypothetical targets. In J. Alspector, R. Goodman, and T. X. Brown (eds.), *Proceedings of the International Workshop on Applications of Neural Networks to Telecommunications 2*, pp. 300–307. Erlbaum, Hillsdale, NJ.
- Gardner, M. (1973). Mathematical games. *Scientific American*, 228(1):108–115.
- Geibel, P. (2001). Reinforcement learning with bounded risk. In *Proceedings of the 18th International Conference on Machine Learning*. (cit. on p. 24).
- Geist, M., Scherrer, B. (2014). Off-policy learning with eligibility traces: A survey. *Journal of Machine Learning Research* 15:289–333.
- Gelperin, A., Hopfield, J. J., Tank, D. W. (1985). The logic of *limax* learning. In A. Selverston (ed.), *Model Neural Networks and Behavior*, pp. 247–261. Plenum Press, New York. (cit. on p. 22).
- Genesereth, M., Thielscher, M. (2014). General game playing. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(2), 1–229.
- Gershman, S. J., Moustafa, A. A., and Ludvig, E. A. (2013). Time representation in reinforcement learning models of the basal ganglia. *Frontiers in computational neuroscience*, 7.
- Gershman, S. J. and Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinions in Neurobiology*, 20:251–256.
- Gershman, S. J., Pesaran, B., and Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience* 29(43):13524–13531.
- Ghiasian, S., Rafiee, B., Sutton, R. S. (2016). A first empirical study of emphatic temporal difference learning. Workshop on Continual Learning and Deep Learning at the Conference on Neural Information Processing Systems. ArXiv:1705.04185.
- Gibbs, C. M., Cool, V., Land, T., Kehoe, E. J., and Gormezano, I. (1991). Second-order conditioning of the rabbits nictitating membrane response. *Integrative Physiological and Behavioral Science* 26(4):282–295.
- Gittins, J. C., Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani, K. Sarkadi, and I. Vincze (eds.), *Progress in Statistics*, pp. 241–266. North-Holland, Amsterdam-London.
- Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. MIT Press, Cambridge, MA.
- Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654.

- Glimcher, P. W. and Fehr E., editors (2013). *Neuroeconomics: Decision making and the brain, Second Edition*. Academic Press.
- Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA.
- Goldstein, H. (1957). *Classical Mechanics*. Addison-Wesley, Reading, MA.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Goodwin, G. C., Sin, K. S. (1984). *Adaptive Filtering Prediction and Control*. Prentice-Hall, Englewood Cliffs, NJ.
- Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111(1):3–32.
- Gordon, G. J. (1995). Stable function approximation in dynamic programming. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 261–268. Morgan Kaufmann, San Francisco. An expanded version was published as Technical Report CMU-CS-95-103. Carnegie Mellon University, Pittsburgh, PA, 1995.
- Gordon, G. J. (1996a). Chattering in SARSA(λ). CMU learning lab internal report.
- Gordon, G. J. (1996b). Stable fitted reinforcement learning. In D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1052–1058. MIT Press, Cambridge, MA.
- Gordon, G. J. (1999). *Approximate solutions to Markov decision processes*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.
- Gordon, G. J. (2001). Reinforcement learning with function approximation converges to a region. *Advances in neural information processing systems*.
- Graybiel, A. M. (2000). The basal ganglia. *Current Biology*, 10(14):R509–R511.
- Greensmith, E., Bartlett, P. L., Baxter, J. (2001). Variance reduction techniques for gradient estimates in reinforcement learning. In *Advances in Neural Information Processing Systems: Proceedings of the 2000 Conference*, pp. 1507–1514.
- Greensmith, E., Bartlett, P. L., Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research* 5(Nov), 1471–1530.
- Griffith, A. K. (1966). A new machine learning technique applied to the game of checkers. Technical Report Project MAC, Artificial Intelligence Memo 94. Massachusetts Institute of Technology, Cambridge, MA.
- Griffith, A. K. (1974). A comparison and evaluation of three machine learning procedures as applied to the game of checkers. *Artificial Intelligence*, 5:137–148.
- Grondman, I., Busoniu, L., Lopes, G. A., Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42(6), 1291–1307.
- Grossberg, S. (1975). A neural model of attention, reinforcement, and discrimination learning. *International Review of Neurobiology*, 18:263–327.
- Grossberg, S. and Schmajuk, N. A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks*, 2(2):79–102.
- Gullapalli, V. (1990). A stochastic reinforcement algorithm for learning real-valued functions. *Neural Networks*, 3:671–692.

- Gullapalli, V. and Barto, A. G. (1992). Shaping as a method for accelerating reinforcement learning. In *Proceedings of the 1992 IEEE International Symposium on Intelligent Control*, pages 554–559. IEEE.
- Gurney, K., Prescott, T. J., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia I. A new functional anatomy. *Biological cybernetics*, 84(6):401–410.
- Gurvits, L., Lin, L.-J., Hanson, S. J. (1994). Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems. Preprint.
- Hackman, L. (2012). *Faster Gradient-TD Algorithms* (MSc dissertation, University of Alberta).
- Hallak, A., Tamar, A., Munos, R., Mannor, S. (2016). Generalized emphatic temporal difference learning: Bias-variance analysis. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- Hammer, M. (1997). The neural basis of associative reward learning in honeybees. *Trends in Neuroscience*, 20:245–252.
- Hammer, M. and Menzel, R. (1995). Learning and memory in the honeybee. *Journal of Neuroscience*, 15(3):1617–1630.
- Hampson, S. E. (1983). *A Neural Model of Adaptive Behavior*. Ph.D. thesis, University of California, Irvine.
- Hampson, S. E. (1989). *Connectionist Problem Solving: Computational Aspects of Biological Learning*. Birkhauser, Boston.
- Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, 28(22):5623–5630.
- Hassabis, D. and Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in cognitive sciences*, 11(7):299–306.
- Hasselt, H. van (2010). Double Q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621.
- Hasselt, H. van (2011). *Insights in Reinforcement Learning: Formal Analysis and Empirical Evaluation of Temporal-difference Learning*. SIKS dissertation series number 2011-04.
- Hasselt, H. van (2012). Reinforcement learning in continuous state and action spaces. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 207–251. Springer Berlin Heidelberg.
- Hasselt, H. van and Sutton, R. S. (2015). Learning to predict independent of span. ArXiv 1508.04582.
- Hawkins, R. D., Kandel, E. R. (1984). Is there a cell-biological alphabet for simple forms of learning? *Psychological Review*, 91:375–391. (cit. on p. 22).
- He, K., Huertas, M., Hong, S. Z., Tie, X., Hell, J. W., Shouval, H., and Kirkwood, A. (2015). Distinct eligibility traces for LTP and LTD in cortical synapses. *Neuron*, 88(3):528–538.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 1992 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. John Wiley and Sons Inc., New York. Reissued by Lawrence Erlbaum Associates Inc., Mahwah NJ, 2002.

- Heger, M. (1994). Consideration of risk in reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pp. 105–111. Morgan Kaufmann. (cit. on p. 24).
- Hengst, B. (2012). Hierarchical approaches. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 293–323. Springer Berlin Heidelberg.
- Herrnstein, R. J. (1970). On the Law of Effect. *Journal of the Experimental Analysis of Behavior* 13(2), 243–266. (cit. on p. 17).
- Hersh, R., Griego, R. J. (1969). Brownian motion and potential theory. *Scientific American*, 220:66–74.
- Hester, T. and Stone, P. (2012). Learning and using models. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 111–141. Springer Berlin Heidelberg.
- Hesterberg, T. C. (1988). *Advances in importance sampling*, Ph.D. Dissertation, Statistics Department, Stanford University.
- Hilgard, E. R. (1956). *Theories of Learning, Second Edition*. Appleton-Century-Crofts, Inc., New York.
- Hilgard, E. R., Bower, G. H. (1975). *Theories of Learning*. Prentice-Hall, Englewood Cliffs, NJ. (cit. on p. 17).
- Hinton, G. E. (1984). Distributed representations. Technical Report CMU-CS-84-157. Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA.
- Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554.
- Hochreiter, S., Schmidhuber, J. (1997). LSTMs can solve hard time lag problems. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 473–479. MIT Press, Cambridge, MA.
- Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor. (cit. on p. 21).
- Holland, J. H. (1976). Adaptation. In R. Rosen and F. M. Snell (eds.), *Progress in Theoretical Biology*, vol. 4, pp. 263–293. Academic Press, New York.
- Holland, J. H. (1986). Escaping brittleness: The possibility of general-purpose learning algorithms applied to rule-based systems. In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), *Machine Learning: An Artificial Intelligence Approach*, vol. 2, pp. 593–623. Morgan Kaufmann, San Mateo, CA. (cit. on p. 22).
- Hollerman, J. R. and Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1:304–309.
- Houk, J. C., Adams, J. L., Barto, A. G. (1995). A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser (eds.), *Models of Information Processing in the Basal Ganglia*, pp. 249–270. MIT Press, Cambridge, MA. (cit. on p. 23).
- Howard, R. (1960). *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA. (cit. on p. 16).
- Hull, C. L. (1932). The goal-gradient hypothesis and maze learning. *Psychological Review*, 39(1):25–43.
- Hull, C. L. (1943). *Principles of Behavior*. Appleton-Century, New York. (cit. on p. 17).
- Hull, C. L. (1952). *A Behavior System*. Wiley, New York.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167.

- İpek, E., Muthu, O., Martínez, J. F., and Caruana, R. (2008). Self-optimizing memory controllers: A reinforcement learning approach. In *35th International Symposium on Computer Architecture, ISCA'08*, pages 39–50. IEEE.
- Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral cortex*, 17(10):2443–2452.
- Jaakkola, T., Jordan, M. I., Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201.
- Jaakkola, T., Singh, S. P., Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable Markov decision problems. In G. Tesauro, D. S. Touretzky, T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 345–352. MIT Press, Cambridge, MA.
- Joel, D., Niv, Y., and Ruppin, E. (2002). Actor-critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks*, 15(4):535–547.
- Johanson, E. B., Killeen, P. R., Russell, V. A., Tripp, G., Wickens, J. R., Tannock, R., Williams, J., and Sagvolden, T. (2009). Origins of altered reinforcement effects in ADHD. *Behavioral and Brain Functions*, 5(7).
- Johnson, A. and Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of neuroscience*, 27(45):12176–12189.
- Kaelbling, L. P. (1993a). Hierarchical learning in stochastic domains: Preliminary results. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 167–173. Morgan Kaufmann, San Mateo, CA. (cit. on p. 23).
- Kaelbling, L. P. (1993b). *Learning in Embedded Systems*. MIT Press, Cambridge, MA.
- Kaelbling, L. P. (Ed.) (1996). Special triple issue on reinforcement learning, *Machine Learning* 22(1/2/3). (cit. on p. 23).
- Kaelbling, L. P., Littman, M. L., Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285. (cit. on p. 23).
- Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 47:263–291.
- Kakade, S. (2002). A natural policy gradient. *Advances in neural information processing systems* 2, 1531–1538.
- Kakade, S. M. (2003). On the Sample Complexity of Reinforcement Learning (Doctoral dissertation, University of London).
- Kakutani, S. (1945). Markov processes and the Dirichlet problem. *Proceedings of the Japan Academy*, 21:227–233.
- Kalos, M. H., Whitlock, P. A. (1986). *Monte Carlo Methods*. Wiley, New York.
- Kamin, L. J. (1968). “Attention-like” processes in classical conditioning. In Jones, M. R., editor, *Miami Symposium on the Prediction of Behavior, 1967: Aversive Stimulation*, pages 9–31. University of Miami Press, Coral Gables, Florida.
- Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In Campbell, B. A. and Church, R. M., editors, *Punishment and Aversive Behavior*, pages 279–296. Appleton-Century-Crofts, New York, NY.
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth A. J., editors (2013). *Principles of Neural Science, Fifth Edition*. McGraw-Hill Companies, Inc.
- Kanerva, P. (1988). *Sparse Distributed Memory*. MIT Press, Cambridge, MA.

- Kanerva, P. (1993). Sparse distributed memory and related models. In M. H. Hassoun (ed.), *Associative Neural Memories: Theory and Implementation*, pp. 50–76. Oxford University Press, New York.
- Karmpatzakis, N. and Langford, J. (2010). Online importance weight aware updates. ArXiv:1011.1576.
- Kashyap, R. L., Blaydon, C. C., Fu, K. S. (1970). Stochastic approximation. In J. M. Mendel and K. S. Fu (eds.), *Adaptive, Learning, and Pattern Recognition Systems: Theory and Applications*, pp. 329–355. Academic Press, New York.
- Kearns, M., Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3), 209–232.
- Keerthi, S. S., Ravindran, B. (1997). Reinforcement learning. In E. Fiesler and R. Beale (eds.), *Handbook of Neural Computation*, C3. Oxford University Press, New York. (cit. on p. 23).
- Kehoe, E. J. (1982). Conditioning with serial compound stimuli: Theoretical and empirical issues. *Experimental Animal Behavior*, 1:30–65.
- Kehoe, E. J., Schreurs, B. G., and Graham, P. (1987). Temporal primacy overrides prior training in serial compound conditioning of the rabbits nictitating membrane response. *Animal Learning & Behavior*, 15(4):455–464.
- Keiflin, R. and Janak, P. H. (2015). Dopamine prediction errors in reward learning and addiction: From theory to neural circuitry. *Neuron*, 88(2):247–263.
- Kimble, G. A. (1961). *Hilgard and Marquis' Conditioning and Learning*. Appleton-Century-Crofts, New York. (cit. on p. 17).
- Kimble, G. A. (1967). *Foundations of Conditioning and Learning*. Appleton-Century-Crofts, New York. (cit. on p. 17).
- Klopf, A. H. (1972). Brain function and adaptive systems—A heterostatic theory. Technical Report AFCRL-72-0164, Air Force Cambridge Research Laboratories, Bedford, MA. A summary appears in *Proceedings of the International Conference on Systems, Man, and Cybernetics*. IEEE Systems, Man, and Cybernetics Society, Dallas, TX, 1974. (cit. on pp. 21, 22).
- Klopf, A. H. (1975). A comparison of natural and artificial intelligence. *SIGART Newsletter*, 53:11–13. (cit. on p. 21).
- Klopf, A. H. (1982). *The Hedonistic Neuron: A Theory of Memory, Learning, and Intelligence*. Hemisphere, Washington, DC. (cit. on p. 21).
- Klopf, A. H. (1988). A neuronal model of classical conditioning. *Psychobiology*, 16:85–125. (cit. on p. 22).
- Kober, J. and Peters, J. (2012). Reinforcement learning in robotics: A survey. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State-of-the-Art*, pages 579–610. Springer-Verlag, Berlin.
- Kocsis, L., Szepesvári, Cs. (2006). Bandit based Monte-Carlo planning. In *Proceedings of the European Conference on Machine Learning*, 282–293. Springer Berlin Heidelberg.
- Kohonen, T. (1977). *Associative Memory: A System Theoretic Approach*. Springer-Verlag, Berlin.
- Koller, D., Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- Kolodziejksi, C., Porr, B., and Wörgötter, F. (2009). On the asymptotic equivalence between differential Hebbian and temporal difference learning. *Neural computation*, 21(4):1173–1202.

- Kolter, J. Z. (2011). The fixed points of off-policy TD. *Advances in Neural Information Processing Systems 24*, pp. 2169–2177.
- Konidaris, G. D., Osentoski, S., Thomas, P. S. (2011). Value function approximation in reinforcement learning using the Fourier basis, *Proceedings of the Twenty-Fifth Conference of the Association for the Advancement of Artificial Intelligence*, pp. 380–385.
- Korf, R. E. (1988). Optimal path finding algorithms. In L. N. Kanal and V. Kumar (eds.), *Search in Artificial Intelligence*, pp. 223–267. Springer Verlag, Berlin.
- Korf, R. E. (1990). Real-time heuristic search. *Artificial Intelligence* 42(2–3), 189–211.
- Koshland, D. E. (1980). *Bacterial Chemotaxis as a Model Behavioral System*. Raven Press, New York.
- Koza, J. R. (1992). *Genetic programming: On the programming of computers by means of natural selection* (Vol. 1). MIT press. (cit. on p. 21).
- Kraft, L. G., Campagna, D. P. (1990). A summary comparison of CMAC neural network and traditional adaptive control systems. In T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 143–169. MIT Press, Cambridge, MA.
- Kraft, L. G., Miller, W. T., Dietz, D. (1992). Development and application of CMAC neural network-based control. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 215–232. Van Nostrand Reinhold, New York.
- Kumar, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM Journal of Control and Optimization*, 23:329–380.
- Kumar, P. R., Varaiya, P. (1986). *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice-Hall, Englewood Cliffs, NJ.
- Kumar, V., Kanal, L. N. (1988). The CDP: A unifying formulation for heuristic search, dynamic programming, and branch-and-bound. In L. N. Kanal and V. Kumar (eds.), *Search in Artificial Intelligence*, pp. 1–37. Springer-Verlag, Berlin.
- Kushner, H. J., Dupuis, P. (1992). *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, New York.
- Lagoudakis, M., Parr, R. (2003). Least squares policy iteration. *Journal of Machine Learning Research* 4:1107–1149.
- Lai, T. L., Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.
- Lakshminarayanan, S. and Narendra, K. S. (1982). Learning algorithms for two-person zero-sum stochastic games with incomplete information: A unified approach. *SIAM Journal of Control and Optimization*, 20:541–552.
- Lammel, S., Lim, B. K., and Malenka, R. C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology*, 76:353–359.
- Lane, S. H., Handelman, D. A., Gelfand, J. J. (1992). Theory and development of higher-order CMAC neural networks. *IEEE Control Systems* 12(2):23–30.
- Lang, K. J., Waibel, A. H., Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:33–43.
- Lange, S., Gabel, T., and Riedmiller, M. (2012). Batch reinforcement learning. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 45–73. Springer Berlin Heidelberg.

- LeCun, Y. (1985). Une procédure d'apprentissage pour réseau à seuil asymétrique (a learning scheme for asymmetric threshold networks). In *Proceedings of Cognitiva 85*, Paris, France.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Legenstein, R., Pecevski, D., and Maass, W. (2008). A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS Computational Biology*, 4(10).
- Levy, W. B. and Steward, D. (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, 8:791–797.
- Lewis, F. L., Liu, D. (Eds.). (2012). *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*. John Wiley and Sons. (cit. on p. 23).
- Lewis, R. L., Howes, A., and Singh, S. (2014). Computational rationality: Linking mechanism and behavior through utility maximization. *Topics in Cognitive Science*, 6(2):279–311.
- Li, L. (2012). Sample complexity bounds of exploration. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 175–204. Springer Berlin Heidelberg.
- Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670. ACM.
- Lin, C.-S., Kim, H. (1991). CMAC-based adaptive critic self-learning control. *IEEE Transactions on Neural Networks*, 2:530–533.
- Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8:293–321.
- Lin, L.-J., Mitchell, T. (1992). Reinforcement learning with hidden states. In *Proceedings of the Second International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pp. 271–280. MIT Press, Cambridge, MA.
- Littman, M. L. (1994). Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 157–163. Morgan Kaufmann, San Francisco.
- Littman, M. L., Cassandra, A. R., Kaelbling, L. P. (1995). Learning policies for partially observable environments: Scaling up. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 362–370. Morgan Kaufmann, San Francisco.
- Littman, M. L., Dean, T. L., Kaelbling, L. P. (1995). On the complexity of solving Markov decision problems. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pp. 394–402.
- Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Berlin, Springer-Verlag.
- Liu, W., Pokharel, P. P., and Principe, J. C. (2008). The kernel least-mean-square algorithm. *IEEE Transactions on Signal Processing* 56(2):543–554.
- Ljung, L. (1998). System identification. In Procházka, A., Uhlíř, J., Rayner, P. W. J., and Kingsbury, N. G., editors, *Signal Analysis and Prediction*, pages 163–173. Springer Science + Business Media New York, LLC.
- Ljung, L., Söderstrom, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.
- Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1):145–163.

- Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28:47–66. (cit. on p. 16).
- Luce, D. (1959). *Individual Choice Behavior*. Wiley, New York.
- Ludvig, E. A., Bellemare, M. G., and Pearson, K. G. (2011). A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives. In Alonso, E. and Mondragón, E., editors, *Computational neuroscience for advancing artificial intelligence: Models, methods and applications*, pages 111–44. Medical Information Science Reference, Hershey PA.
- Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, 20(12):3034–3054.
- Ludvig, E. A., Sutton, R. S., and Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & behavior*, 40(3):305–319.
- Machado, A. (1997). Learning the temporal dynamics of behavior. *Psychological review*, 104(2):241–265.
- Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4):276–298.
- Mackintosh, N. J. (1983). *Conditioning and Associative Learning*. Oxford: Clarendon Press.
- Maclin, R. and Shavlik, J. W. (1994). Incorporating advice into agents that learn from reinforcements. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 694–699. AAAI Press, Menlo Park, CA.
- Maei, H. R. (2011). *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta.
- Maei, H. R. and Sutton, R. S. (2010). GQ(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence*, pp. 91–96.
- Maei, H. R., Szepesvári, Cs., Bhatnagar, S., Precup, D., Silver, D., and Sutton, R. S. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, pp. 1204–1212.
- Maei, H. R., Szepesvári, Cs., Bhatnagar, S., and Sutton, R. S. (2010). Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 719–726.
- Mahadevan, S. (1996). Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22:159–196.
- Mahadevan, S. and Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55:311–365.
- Mahadevan, S., Liu, B., Thomas, P., Dabney, W., Giguere, S., Jacek, N., Gemp, I., Liu, J. (2014). Proximal reinforcement learning: A new theory of sequential decision making in primal-dual spaces. ArXiv preprint arXiv:1405.6757.
- Mahmood, A. R. (2017). Incremental Off-policy Reinforcement Learning Algorithms. University of Alberta PhD thesis.
- Mahmood, A. R. and Sutton, R. S. (2015). Off-policy learning based on weighted importance sampling with linear computational complexity. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, Amsterdam, Netherlands.
- Mahmood, A. R., Sutton, R. S., Degris, T., and Pilarski, P. M. (2012). Tuning-free step-size adaptation. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE*

- International Conference on Acoustics, Speech and Signal Processing* (pp. 2121–2124). IEEE.
- Mahmood, A. R., van Hasselt, H., and Sutton, R. S. (2014). Weighted importance sampling for off-policy learning with linear function approximation. *Advances in Neural Information Processing Systems 27*.
- Mahmood, A. R., Yu, H., Sutton, R. S. (2017). Multi-step off-policy learning without importance sampling ratios. ArXiv 1702.03006.
- Marbach, P., Tsitsiklis, J. N. (2001). Simulation-based optimization of Markov reward processes. *IEEE Transactions on Automatic Control* 46(2), 191–209. Also MIT Technical Report LIDS-P-2411 (1998).
- Markey, K. L. (1994). Efficient learning of multiple degree-of-freedom control problems with quasi-independent Q-agents. In M. C. Mozer, P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend (eds.), *Proceedings of the 1990 Connectionist Models Summer School*. Erlbaum, Hillsdale, NJ.
- Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science*, 275:213–215.
- Martínez, J. F. and İpek, E. (2009). Dynamic multicore resource management: A machine learning approach. *Micro, IEEE*, 29(5):8–17.
- Mataric, M. J. (1994). Reward functions for accelerated learning. In *Machine Learning: Proceedings of the Eleventh international conference*, pages 181–189.
- Matsuda, W., Furuta, T., Nakamura, K. C., Hioki, H., Fujiyama, F., Arai, R., and Kaneko, T. (2009). Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. *The Journal of Neuroscience*, 29(2):444–453.
- Mazur, J. E. (1994). *Learning and Behavior*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ. (cit. on p. 17).
- McCallum, A. K. (1993). Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 190–196. Morgan Kaufmann, San Mateo, CA.
- McCallum, A. K. (1995). *Reinforcement Learning with Selective Perception and Hidden State*. Ph.D. thesis, University of Rochester, Rochester, NY.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5(4):115–133.
- Melo, F. S., Meyn, S. P., Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning* (pp. 664–671).
- Mendel, J. M. (1966). A survey of learning control systems. *ISA Transactions*, 5:297–303. (cit. on p. 19).
- Mendel, J. M., McLaren, R. W. (1970). Reinforcement learning control and pattern recognition systems. In J. M. Mendel and K. S. Fu (eds.), *Adaptive, Learning and Pattern Recognition Systems: Theory and Applications*, pp. 287–318. Academic Press, New York. (cit. on p. 19).
- Michie, D. (1961). Trial and error. In S. A. Barnett and A. McLaren (eds.), *Science Survey, Part 2*, pp. 129–145. Penguin, Harmondsworth. (cit. on p. 19).
- Michie, D. (1963). Experiments on the mechanisation of game learning. 1. characterization of the model and its parameters. *Computer Journal*, 1:232–263. (cit. on p. 19).

- Michie, D. (1974). *On Machine Intelligence*. Edinburgh University Press, Edinburgh. (cit. on p. 20).
- Michie, D., Chambers, R. A. (1968). BOXES: An experiment in adaptive control. In E. Dale and D. Michie (eds.), *Machine Intelligence 2*, pp. 137–152. Oliver and Boyd, Edinburgh. (cit. on p. 19).
- Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290 (cit. on p. 24).
- Miller, R. (1981). *Meaning and Purpose in the Intact Brain: A Philosophical, Psychological, and Biological Account of Conscious Process*. Clarendon Press, Oxford.
- Miller, S., Williams, R. J. (1992). Learning to control a bioreactor using a neural net Dyna-Q system. In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, pp. 167–172. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.
- Miller, W. T., An, E., Glanz, F., Carter, M. (1990). The design of CMAC neural networks for control. *Adaptive and Learning Systems* 1:140–145.
- Miller, W. T., Glanz, F. H. (1996). *UNH-CMAC version 2.1: The University of New Hampshire Implementation of the Cerebellar Model Arithmetic Computer - CMAC*. Robotics Laboratory Technical Report, University of New Hampshire, Durham, New Hampshire.
- Miller, W. T., Scalera, S. M., Kim, A. (1994). Neural network control of dynamic balance for a biped walking robot. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, pp. 156–161. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.
- Minsky, M. L. (1954). *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem*. Ph.D. thesis, Princeton University. (cit. on pp. 18, 21).
- Minsky, M. L. (1961). Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49:8–30. Reprinted in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 406–450. McGraw-Hill, New York, 1963. (cit. on pp. 19, 22).
- Minsky, M. L. (1967). *Computation: Finite and Infinite Machines*. Prentice-Hall, Englewood Cliffs, NJ.
- 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., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.
- Modayil, J. and Sutton, R. S. (2014). Prediction driven behavior: Learning predictions that drive fixed responses. In *AAAI-14 Workshop on Artificial Intelligence and Robotics*, Quebec City, Canada.
- Modayil, J., White, A., and Sutton, R. S. (2014). Multi-timescale nexting in a reinforcement learning robot. *Adaptive Behavior*, 22(2):146–160.
- Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., and Sejnowski, T. J. (1992). Using aperiodic reinforcement for directed self-organization during development. In *Advances in neural information processing systems 5*, pages 969–976.
- Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377(6551):725–728.

- Montague, P. R., Dayan, P., Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16:1936–1947. (cit. on p. 23).
- Montague, P. R., Dolan, R. J., Friston, K. J., and Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences* 16(1):72–80.
- Montague, P. R. and Sejnowski, T. J. (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learning mechanisms. *Learning & Memory*, 1:1–33.
- Moore, A. W. (1990). *Efficient Memory-Based Learning for Robot Control*. Ph.D. thesis, University of Cambridge.
- Moore, A. W. (1994). The parti-game algorithm for variable resolution reinforcement learning in multidimensional spaces. In J. D. Cohen, G. Tesauro and J. Alspector (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1993 Conference*, pp. 711–718. Morgan Kaufmann, San Francisco.
- Moore, A. W., Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13:103–130.
- Moore, A. W., Schneider, J., and Deng, K. (1997). Efficient locally weighted polynomial regression predictions. In *Proceedings of the 1997 International Machine Learning Conference*. Morgan Kaufmann.
- Moore, J. W. and Blazis, D. E. J. (1989). Simulation of a classically conditioned response: A cerebellar implementation of the sutton-barto-desmond model. In Byrne, J. H. and Berry, W. O., editors, *Neural Models of Plasticity*, pages 187–207. Academic Press, San Diego, CA.
- Moore, J. W., Choi, J.-S., and Brunzell, D. H. (1998). Predictive timing under temporal uncertainty: The time derivative model of the conditioned response. In Rosenbaum, D. A. and Collyer, C. E., editors, *Timing of Behavior*, pages 3–34. MIT Press, Cambridge, MA.
- Moore, J. W., Desmond, J. E., Berthier, N. E., Blazis, E. J., Sutton, R. S., and Barto, A. G. (1986). Simulation of the classically conditioned nictitating membrane response by a neuron-like adaptive element: I. Response topography, neuronal firing, and interstimulus intervals. *Behavioural Brain Research*, 21:143–154. (cit. on p. 22).
- Moore, J. W., Marks, J. S., Castagna, V. E., and Polewan, R. J. (2001). Parameter stability in the TD model of complex CR topographies. Society for Neuroscience Abstract 642.2.
- Moore, J. W. and Schmajuk, N. A. (2008). Kamin blocking. *Scholarpedia*, 3(5):3542.
- Moore, J. W. and Stickney, K. J. (1980). Formation of attentional-associative networks in real time: Role of the hippocampus and implications for conditioning. *Physiological Psychology*, 8(2):207–217.
- Mukundan, J. and Martínez, J. F. (2012). MORSE: Multi-objective reconfigurable self-optimizing memory scheduler. In IEEE 18th International Symposium on High Performance Computer Architecture (HPCA), pages 1–12.
- Müller, M. (2002). Computer Go. *Artificial Intelligence*, 134(1):145–179.
- Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. (2016). Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1046–1054.
- Naddaf, Y. (2010). *Game-independent AI agents for playing Atari 2600 console games*. PhD thesis, University of Alberta.
- Narendra, K. S., Thathachar, M. A. L. (1974). Learning automata—A survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 4:323–334. (cit. on p. 20).

- Narendra, K. S., Thathachar, M. A. L. (1989). *Learning Automata: An Introduction*. Prentice-Hall, Englewood Cliffs, NJ. (cit. on p. 20).
- Narendra, K. S. and Wheeler, R. M. (1983). An n -player sequential stochastic game with identical payoffs. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:1154–1158.
- Narendra, K. S., Wheeler, R. M. (1986). Decentralized learning in finite Markov chains. *IEEE Transactions on Automatic Control*, AC31(6):519–526.
- Nedić, A., Bertsekas, D. P. (2003). Least squares policy evaluation algorithms with linear function approximation. *Discrete Event Dynamic Systems* 13(1–2):79–110.
- Ng, A. Y. (2003). *Shaping and policy search in reinforcement learning*. PhD thesis, University of California, Berkeley, Berkeley, CA.
- Ng, A. Y., Harada, D., and Russell, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In Bratko, I. and Dzeroski, S., editors, *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999)*, volume 99, pp. 278–287.
- Ng, A. Y. and Russell, S. J. (2000). Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pp. 663–670.
- Nie, J., Haykin, S. (1996). A dynamic channel assignment policy through Q-learning. CRL Report 334. Communications Research Laboratory, McMaster University, Hamilton, Ontario.
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154.
- Niv, Y., Daw, N. D., and Dayan, P. (2005). How fast to work: Response vigor, motivation and tonic dopamine. In Yeiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 1019–1026. MIT Press, Cambridge, MA.
- Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.
- Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, 10(8):375–381.
- Nowé, A., Vrancx, P., and Hauwere, Y.-M. D. (2012). Game theory and multi-agent reinforcement learning. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State-of-the-Art*, pages 441–467. Springer-Verlag, Berlin. (cit. on p. 21).
- Nutt, D. J., Lingford-Hughes, A., Erritzoe, D., and Stokes, P. R. A. (2015). The dopamine theory of addiction: 40 years of highs and lows. *Nature Reviews Neuroscience*, 16:305–312.
- O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329–337.
- O'Doherty, J. P., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.
- Oh, J., Guo, X., Lee, H., Lewis, R. L., and Singh, S. (2015). Action-conditional video prediction using deep networks in Atari games. In *Advances in Neural Information Processing Systems*, pages 2845–2853.
- Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., and Spiers, H. J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. *Elife*, 4:e06063.

- Olds, J. and Milner, P. (1954). Positive reinforcement produced by electrical stimulation of the septal area and other regions of rat brain. *Journal of Comparative and Physiological Psychology*, 47(6):419–427.
- Oliehoek, F. A. (2012). Decentralized POMDPs. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 471–503. Springer Berlin Heidelberg.
- Omohundro, S. M. (1987). Efficient algorithms with neural network behavior. Technical Report, Department of Computer Science, University of Illinois at Urbana-Champaign.
- O'Reilly, R. C. and Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328.
- O'Reilly, R. C., Frank, M. J., Hazy, T. E., and Watz, B. (2007). PVLV: the primary value and learned value Pavlovian learning algorithm. *Behavioral neuroscience*, 121(1):31–49.
- Orenstein, J. A. (1982). Multidimensional tries used for associative searching. *Information Processing Letters* 14(4):150–157.
- Ormoneit, D. and Sen, Š. (2002). Kernel-based reinforcement learning. *Machine learning* 49(2–3):161–178.
- Otterlo, M. van (2009). *The Logic of Adaptive Behavior*. IOS Press.
- Otterlo, M. van (2012). Solving relational and first-order logical markov decision processes: A survey. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 253–292. Springer Berlin Heidelberg.
- Oudeyer, P.-Y. and Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1.
- Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2):265–286.
- Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature* 441(7090):223–226.
- Page, C. V. (1977). Heuristics for signature table analysis as a pattern recognition technique. *IEEE Transactions on Systems, Man, and Cybernetics*, 7:77–86.
- Pagnoni, G., Zink, C. F., Montague, P. R., and Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature neuroscience*, 5(2):97–98.
- Pan, W.-X., Schmidt, R., Wickens, J. R., and Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: Evidence for eligibility traces in the reward-learning network. *The Journal of Neuroscience*, 25(26):6235–6242.
- Park, J., Kim, J., Kang, D. (2005). An RLS-based natural actor-critic algorithm for locomotion of a two-linked robot arm. *Computational Intelligence and Security*, 65–72.
- Parker, D. B. (1985). *Learning Logic*. Technical Report TR-47, Center for Computational Research in Economics and Management Science, MIT, Cambridge, MA.
- Parks, P. C., Militzer, J. (1991). Improved allocation of weights for associative memory storage in learning control systems. *IFAC Design Methods of Control Systems*, Zurich, Switzerland, 507–512.
- Parr, R., Russell, S. (1995). Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1088–1094. Morgan Kaufmann.
- Pavlov, P. I. (1927). *Conditioned Reflexes*. Oxford University Press, London.

- Pawlak, V. and Kerr, J. N. D. (2008). Dopamine receptor activation is required for corticostratial spike-timing-dependent plasticity. *The Journal of Neuroscience*, 28(10):2435–2446.
- Pawlak, V., Wickens, J. R., Kirkwood, A., and Kerr, J. N. D. (2010). Timing is not everything: neuromodulation opens the STDP gate. *Frontiers in synaptic neuroscience*, 2.
- Pearce, J. M. and Hall, G. (1980). A model for Pavlovian learning: Variation in the effectiveness of conditioning but not unconditioned stimuli. *Psychological Review*, 87(6):532–552.
- Pearl, J. (1984). *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, Reading, MA.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4), 669–688.
- Pecevski, D., Maass, W., and Legenstein, R. A. (2007). Theoretical analysis of learning with reward-modulated spike-timing-dependent plasticity. In *Advances in Neural Information Processing Systems*, pp. 881–888.
- Peng, J. (1993). *Efficient Dynamic Programming-Based Learning for Control*. Ph.D. thesis, Northeastern University, Boston.
- Peng, J. (1995). Efficient memory-based dynamic programming. In *12th International Conference on Machine Learning*, pp. 438–446.
- Peng, J., Williams, R. J. (1993). Efficient learning and planning within the Dyna framework. *Adaptive Behavior*, 1(4):437–454.
- Peng, J., Williams, R. J. (1994). Incremental multi-step Q-learning. In W. W. Cohen and H. Hirsh (eds.), *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 226–232. Morgan Kaufmann, San Francisco.
- Peng, J., Williams, R. J. (1996). Incremental multi-step Q-learning. *Machine Learning*, 22:283–290.
- Perkins, T. J., Pendrith, M. D. (2002). On the existence of fixed points for Q-learning and Sarsa in partially observable domains. In *Proceedings of the International Conference on Machine Learning*, pp. 490–497.
- Perkins, T. J., Precup, D. (2003). A convergent form of approximate policy iteration. In *Advances in neural information processing systems, proceedings of the 2002 conference*, pp. 1595–1602.
- Peters, J. and Büchel, C. (2010). Neural representations of subjective reward value. *Behavioral brain research*, 213(2):135–141.
- Peters, J., Schaal, S. (2008). Natural actor-critic. *Neurocomputing* 71(7), 1180–1190.
- Peters, J., Vijayakumar, S., Schaal, S. (2005). Natural actor-critic. In *European Conference on Machine Learning* (pp. 280–291). Springer Berlin Heidelberg.
- Peterson, G. B. (2004). A day of great illumination: B.F. Skinner's discovery of shaping. *Journal of the Experimental Analysis of Behavior*, 82(3):317–328.
- Pezzulo, G., van der Meer, M. A. A., Lansink, C. S., and Pennartz, C. M. A. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Science*, 18(12):647–657.
- Pfeiffer, B. E. and Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497(7447):74–79.
- Phansalkar, V. V., Thathachar, M. A. L. (1995). Local and global optimization algorithms for generalized learning automata. *Neural Computation*, 7:950–973.

- Poggio, T., Girosi, F. (1989). A theory of networks for approximation and learning. A.I. Memo 1140. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.
- Poggio, T., Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.
- Polyak, B. T. (1990). New stochastic approximation type procedures. *Automat. i Telemekh* 7(98–107), 2 (in Russian).
- Polyak, B. T., Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30(4), 838–855.
- Powell, M. J. D. (1987). Radial basis functions for multivariate interpolation: A review. In J. C. Mason and M. G. Cox (eds.), *Algorithms for Approximation*, pp. 143–167. Clarendon Press, Oxford.
- Powell, W. B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, Second edition. John Wiley and Sons. (cit. on p. 23).
- Powers, W. T. (1973). *Behavior: The Control of Perception*. Aldine de Gruyter, Chicago. 2nd expanded edition 2005.
- Precup, D., Sutton, R. S., Dasgupta, S. (2001). Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*.
- Precup, D., Sutton, R. S., Paduraru, C., Koop, A., and Singh, S. (2005). Off-policy learning with options and recognizers. In *Advances in Neural Processing Systems*, pp. 1097–1104.
- Precup, D., Sutton, R. S., Singh, S. (2000). Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766. Morgan Kaufmann.
- Puterman, M. L. (1994). *Markov Decision Problems*. Wiley, New York. (cit. on p. 17).
- Puterman, M. L., Shin, M. C. (1978). Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, 24:1127–1137.
- Quartz, S., Dayan, P., Montague, P. R., and Sejnowski, T. J. (1992). Expectation learning in the brain using diffuse ascending connections. In *Society for Neuroscience Abstracts*, volume 18, page 1210.
- Randløv, J. and Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 463–471.
- Rangel, A., Camerer, C., and Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, 9(7):545–556.
- Rangel, A. and Hare, T. (2010). Neural computations associated with goal-directed choice. *Current opinion in neurobiology*, 20(2):262–270.
- Reddy, G., Celani, A., Sejnowski, T. J., and Vergassola, M. (2016). Learning to soar in turbulent environments. *Proceedings of the National Academy of Sciences*, 113(33):E4877–E4884.
- Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nature Reviews Neuroscience*, 7:967–975.
- Redish, D. A. (2004). Addiction as a computational process gone awry. *Science*, 306(5703):1944–1947.

- Reetz, D. (1977). Approximate solutions of a discounted Markovian decision process. *Bonner Mathematische Schriften*, 98:77–92.
- Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.
- Revusky, S. and Garcia, J. (1970). Learned associations over long delays. In Bower, G., editor, *The psychology of learning and motivation*, volume 4, pages 1–84. Academic Press, Inc., New York.
- Reynolds, J. N. J. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4):507–521.
- Ring, M. B. (1994). *Continual Learning in Reinforcement Environments*. Ph.D. thesis, University of Texas, Austin.
- Ripley, B. D. (2007). *Pattern Recognition and Neural Networks*. Cambridge University Press.
- Rivest, R. L., Schapire, R. E. (1987). Diversity-based inference of finite automata. In *Proceedings of the Twenty-Eighth Annual Symposium on Foundations of Computer Science*, pp. 78–87. Computer Society Press of the IEEE, Washington, DC.
- Rixner, S. (2004). Memory controller optimizations for web servers. In *Proceedings of the 37th annual IEEE/ACM International Symposium on Microarchitecture*, pages 355–366. IEEE Computer Society.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.
- Robertie, B. (1992). Carbon versus silicon: Matching wits with TD-Gammon. *Inside Backgammon*, 2:14–22.
- Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12):1615–1624.
- Romo, R. and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, 63(3):592–624.
- Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington, DC. (cit. on p. 18).
- Ross, S. (1983). *Introduction to Stochastic Dynamic Programming*. Academic Press, New York. (cit. on p. 17).
- Ross, T. (1933). Machines that think. *Scientific American*, pages 206–208. (cit. on p. 18).
- Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. Wiley, New York.
- Rumelhart, D. E., Hinton, G. E., Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. I, *Foundations*. Bradford/MIT Press, Cambridge, MA.
- Rummery, G. A. (1995). *Problem Solving with Reinforcement Learning*. Ph.D. thesis, Cambridge University.
- Rummery, G. A., Niranjan, M. (1994). On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166. Engineering Department, Cambridge University.

- Ruppert, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Cornell University Operations Research and Industrial Engineering Technical Report No. 781.
- Russell, S., Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice-Hall, Englewood Cliffs, NJ. (cit. on p. 24).
- Rust, J. (1996). Numerical dynamic programming in economics. In H. Amman, D. Kendrick, and J. Rust (eds.), *Handbook of Computational Economics*, pp. 614–722. Elsevier, Amsterdam. (cit. on p. 16).
- Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54–67.
- Saddoris, M. P., Cacciapaglia, F., Wightman, R. M., and Carelli, R. M. (2015). Differential dopamine release dynamics in the nucleus accumbens core and shell reveal complementary signals for error prediction and incentive motivation. *The Journal of Neuroscience*, 35(33):11572–11582.
- Saksida, L. M., Raymond, S. M., and Touretzky, D. S. (1997). Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems*, 22(3):231–249.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development*, 3:211–229. Reprinted in E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 71–105. McGraw-Hill, New York, 1963. (cit. on p. 22).
- Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. II—Recent progress. *IBM Journal on Research and Development*, 11:601–617.
- Schaal, S. and Atkeson, C. G. (1994). Robot juggling: Implementation of memory-based learning. *IEEE Control Systems* 14(1):57–71.
- Schmajuk, N. A. (2008). Computational models of classical conditioning. *Scholarpedia*, 3(3):1664.
- Schmidhuber, J. (1991a). Adaptive confidence and adaptive curiosity. Technical Report FKI-149-91, Institut für Informatik, Technische Universität München, Arcisstr. 21, 800 München 2, Germany.
- Schmidhuber, J. (1991b). A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA. MIT Press.
- Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In Pezzulo, G., Butz, M. V., Sigaud, O., and Baldassarre, G., editors, *Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems*, pages 48–76. Springer, Berlin.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61, 85–117.
- Schmidhuber, J., Storck, J., and Hochreiter, S. (1994). Reinforcement driven information acquisition in nondeterministic environments. Technical report, Fakultät für Informatik, Technische Universität München, München, Germany.
- Schultz, D. G., Melsa, J. L. (1967). *State Functions and Linear Control Systems*. McGraw-Hill, New York.

- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.
- Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience* 13(3):900–913.
- Schultz, W., Dayan, P., Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1598. (cit. on p. 23).
- Schultz, W. and Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, 63(3):607–624.
- Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., and Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In Houk, Davis, and Beiser (Eds.) *Models of Information Processing in the Basal Ganglia*, pp. 233–248. MIT Press, Cambridge MA.
- Schumaker, L. L. (1976). *Fitting Surfaces to Scattered Data*. University of Texas at Austin, Dept. of Mathematics.
- Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 298–305. Morgan Kaufmann, San Mateo, CA.
- Schweitzer, P. J., Seidmann, A. (1985). Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110:568–582.
- Selfridge, O. G. (1978). Tracking and trailing: Adaptation in movement strategies. Technical report, Bolt Beranek and Newman, Inc. Unpublished report.
- Selfridge, O. G. (1984). Some themes and primitives in ill-defined systems. In Selfridge, O. G., Rissland, E. L., and Arbib, M. A., editors, *Adaptive Control of Ill-Defined Systems*, pages 21–26. Plenum Press, NY. Proceedings of the NATO Advanced Research Institute on Adaptive Control of Ill-defined Systems, NATO Conference Series II, Systems Science, Vol. 16.
- Selfridge, O. J., Sutton, R. S., Barto, A. G. (1985). Training and tracking in robotics. In A. Joshi (ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pp. 670–672. Morgan Kaufmann, San Mateo, CA.
- Seo, H., Barraclough, D., and Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex*, 17(suppl 1):110–117.
- Seung, H. S. (2003). Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6):1063–1073.
- Shah, A. (2012). Psychological and neuroscientific connections with reinforcement learning. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning: State of the Art*, pages 507–537. Springer-Verlag, Berlin.
- Shannon, C. E. (1950). Programming a computer for playing chess. *Philosophical Magazine*, 41:256–275. (cit. on p. 22).
- Shannon, C. E. (1951). Presentation of a maze-solving machine. In Forester, H. V., editor, *Cybernetics. Transactions of the Eighth Conference*, pages 173–180. Josiah Macy Jr. Foundation. (cit. on p. 18).
- Shannon, C. E. (1952). “Theseus” maze-solving mouse. <http://cyberneticzoo.com/mazesolvers/1952--theseus-maze-solving-mouse--claude-shannon-american/>. (cit. on p. 18).

- Shelton, C. R. (2001). *Importance Sampling for Reinforcement Learning with Multiple Objectives*. PhD thesis, Massachusetts Institute of Technology.
- Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 23rd ACM National Conference*, pp. 517–524. ACM.
- Sherman, J., Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix (abstract). *Annals of Mathematical Statistics* 20:621.
- Shewchuk, J., Dean, T. (1990). Towards learning time-varying functions with high input dimensionality. In *Proceedings of the Fifth IEEE International Symposium on Intelligent Control*, pp. 383–388. IEEE Computer Society Press, Los Alamitos, CA.
- Shimansky, Y. P. (2009). Biologically plausible learning in neural networks: a lesson from bacterial chemotaxis. *Biological Cybernetics*, 101(5–6):379–385.
- Si, J., Barto, A., Powell, W., Wunsch, D. (Eds.). (2004). *Handbook of learning and approximate dynamic programming*. John Wiley and Sons. (cit. on p. 23).
- Silver, D. (2009). *Reinforcement learning and simulation based search in the game of Go*. University of Alberta Doctoral dissertation.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.
- Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M. (2014). Deterministic policy gradient algorithms. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (pp. 387–395).
- Şimşek, Ö., Algörta, S., and Kothiyal, A. (2016). Why most decisions are easy in tetris—and perhaps in other sequential decision problems, as well. *Proceedings of 33rd International Conference on Machine Learning*.
- Singh, S. P. (1992a). Reinforcement learning with a hierarchy of abstract models. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 202–207. AAAI/MIT Press, Menlo Park, CA.
- Singh, S. P. (1992b). Scaling reinforcement learning algorithms by learning variable temporal resolution models. In *Proceedings of the Ninth International Machine Learning Conference*, pp. 406–415. Morgan Kaufmann, San Mateo, CA.
- Singh, S. P. (1993). *Learning to Solve Markovian Decision Processes*. Ph.D. thesis, University of Massachusetts, Amherst. Appeared as CMPSCI Technical Report 93-77.
- Singh, S. P. (Ed.) (2002). Special double issue on reinforcement learning, *Machine Learning* 49(2/3). (cit. on p. 23).
- Singh, S. P., Barto, A. G., and Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, pages 1281–1288, Cambridge MA. MIT Press.
- Singh, S. P., Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 974–980. MIT Press, Cambridge, MA.
- Singh, S. P., Jaakkola, T., Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision problems. In W. W. Cohen and H. Hirsch (eds.), *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 284–292. Morgan Kaufmann, San Francisco.

- Singh, S. P., Jaakkola, T., Jordan, M. I. (1995). Reinforcement learning with soft state aggregation. In G. Tesauro, D. S. Touretzky, T. Leen (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1994 Conference*, pp. 359–368. MIT Press, Cambridge, MA.
- Singh, S. P., Lewis, R. L., and Barto, A. G. (2009). Where do rewards come from? In Taatgen, N. and van Rijn, H., editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2601–2606. Cognitive Science Society.
- Singh, S. P., Lewis, R. L., Barto, A. G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):7082. Special issue on Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.
- Singh, S. P., Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22:123–158.
- Sivarajan, K. N., McEliece, R. J., Ketchum, J. W. (1990). Dynamic channel assignment in cellular radio. In *Proceedings of the 40th Vehicular Technology Conference*, pp. 631–637.
- Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century, New York. (cit. on p. 17).
- Skinner, B. F. (1958). Reinforcement today. *American Psychologist*, 13(3):94–99.
- Skinner, B. F. (1981). Selection by consequences. *Science* 213(4507):501–504.
- Smith, K. S. and Greybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, 79(2):361–374.
- Sofge, D. A., White, D. A. (1992). Applied learning: Optimal control for manufacturing. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 259–281. Van Nostrand Reinhold, New York.
- Sorg, J. D. (2011). *The Optimal Reward Problem: Designing Effective Reward for Bounded Agents*. PhD thesis, Computer Science and Engineering, The University of Michigan.
- Sorg, J., Lewis, R. L., and Singh, S. P. (2010). Reward design via online gradient ascent. In *Advances in Neural Information Processing Systems*, pp. 2190–2198.
- Sorg, J., Singh, S., and Lewis, R. (2010). Internal rewards mitigate agent boundedness. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 1007–1014.
- Spaan, M. T. (2012). Partially observable Markov decision processes. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 387–414. Springer Berlin Heidelberg.
- Spence, K. W. (1947). The role of secondary reinforcement in delayed reward learning. *Psychological Review*, 54(1):1–8.
- Spong, M. W. (1994). Swing up control of the acrobot. In *Proceedings of the 1994 IEEE Conference on Robotics and Automation*, pp. 2356–2361. IEEE Computer Society Press, Los Alamitos, CA.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.
- Staddon, J. E. R. (1983). *Behavior and Learning*. Cambridge University Press, Cambridge.
- Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM* 29(12):1213–1228.

- Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., and Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, 16(7):966–973.
- Sterling, P. and Laughlin, S. (2015). *Principles of Neural Design*. MIT Press, Cambridge, MA.
- Storck, J., Hochreiter, S., and Schmidhuber, J. (1995). Reinforcement-driven information acquisition in non-deterministic environments. In *Proceedings of ICANN'95*, Paris, France, volume 2, pages 159–164.
- Sugiyama, M., Hachiya, H., Morimura, T. (2013). *Statistical Reinforcement Learning: Modern Machine Learning Approaches*. Chapman & Hall/CRC. (cit. on p. 23).
- Suri, R. E., Bargas, J., and Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103(1):65–85.
- Suri, R. E. and Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, 121(3):350–354.
- Suri, R. E. and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871–890.
- Sutton, R. S. (1978a). Learning theory support for a single channel theory of the brain. Unpublished report. (cit. on p. 22).
- Sutton, R. S. (1978b). A unified theory of expectation in classical and instrumental conditioning. Bachelors thesis, Stanford University. (cit. on p. 22).
- Sutton, R. S. (1984). *Temporal Credit Assignment in Reinforcement Learning*. Ph.D. thesis, University of Massachusetts, Amherst. (cit. on pp. 20, 22).
- Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44. (cit. on p. 22).
- Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pp. 216–224. Morgan Kaufmann, San Mateo, CA.
- Sutton, R. S. (1991a). Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin*, 2:160–163. ACM Press.
- Sutton, R. S. (1991b). Planning by incremental dynamic programming. In L. A. Birnbaum and G. C. Collins (eds.), *Proceedings of the Eighth International Workshop on Machine Learning*, pp. 353–357. Morgan Kaufmann, San Mateo, CA.
- Sutton, R. S. (Ed.) (1992). *Reinforcement Learning*. Kluwer Academic Press. Reprinting of a special double issue on reinforcement learning, *Machine Learning* 8(3/4). (cit. on p. 23).
- Sutton, R. S. (1995a). TD models: Modeling the world at a mixture of time scales. In A. Prieditis and S. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 531–539. Morgan Kaufmann, San Francisco.
- Sutton, R. S. (1995b). On the virtues of linear learning and trajectory distributions. *Proceedings of the Workshop on Value Function Approximation* at the International Conference on Machine Learning.
- Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1038–1044. MIT Press, Cambridge, MA.

- Sutton, R. S. (2009). The grand challenge of predictive empirical abstract knowledge. *Working Notes of the IJCAI-09 Workshop on Grand Challenges for Reasoning from Experiences*.
- Sutton, R. S. (2015a). Introduction to reinforcement learning with function approximation. Tutorial at the Conference on Neural Information Processing Systems, Montreal, December 7, 2015.
- Sutton, R. S. (2015b). True online Emphatic TD(λ): Quick reference and implementation guide. ArXiv:1507.07147. Code is available in Python and C++ by downloading the source files of this arXiv paper as a zip archive.
- Sutton, R. S. (1978c). Single channel theory: A neuronal theory of learning. *Brain Theory Newsletter*, 4:72–75. Center for Systems Neuroscience, University of Massachusetts, Amherst, MA. (cit. on p. 22).
- Sutton, R. S., Barto, A. G. (1981a). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170. (cit. on p. 22).
- Sutton, R. S., Barto, A. G. (1981b). An adaptive network that constructs and uses an internal model of its world. *Cognition and Brain Theory*, 3:217–246.
- Sutton, R. S., Barto, A. G. (1987). A temporal-difference model of classical conditioning. In *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, pp. 355–378. Erlbaum, Hillsdale, NJ. (cit. on p. 22).
- Sutton, R. S., Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel and J. Moore (eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pp. 497–537. MIT Press, Cambridge, MA. (cit. on p. 22).
- Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, Cs., and Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 993–1000. ACM.
- Sutton, R. S., Maei, H. R., and Szepesvári, Cs. (2009). A convergent $O(d^2)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pp. 1609–1616.
- Sutton, R. S., Mahmood, A. R., Precup, D., van Hasselt, H. (2014). A new Q(λ) with interim forward view and Monte Carlo equivalence. *International Conference on Machine Learning 31. JMLR W&CP 32(2)*.
- Sutton, R. S., Mahmood, A. R., White, M. (2016). An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research* 17(73):1–29.
- Sutton, R. S., McAllester, D. A., Singh, S. P., Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 99*, pp. 1057–1063.
- Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the Tenth International Conference on Autonomous Agents and Multiagent Systems*, pp. 761–768, Taipei, Taiwan.
- Sutton, R. S., Pinette, B. (1985). The learning of world models by connectionist networks. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*, pp. 54–64.

- Sutton, R. S., Singh, S. (1994). On bias and step size in temporal-difference learning. In *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems*, pp. 91–96. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.
- Sutton, R. S., Whitehead, D. S. (1993). Online learning with random representations. In *Proceedings of the Tenth International Machine Learning Conference*, pp. 314–321. Morgan Kaufmann, San Mateo, CA.
- Sutton, R.S., Precup, D., Singh, S. (1999). Between MDPs and semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, 112:181–21.
- Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4(1), 1–103. (cit. on p. 23).
- Szita, I. (2012). Reinforcement learning in games. In *Reinforcement Learning* (pp. 539–577). Springer Berlin Heidelberg. (cit. on p. 21).
- Tadepalli, P., Ok, D. (1994). H-learning: A reinforcement learning method to optimize undiscounted average reward. Technical Report 94-30-01. Oregon State University, Computer Science Department, Corvallis.
- Tadepalli, P. and Ok, D. (1996). Scaling up average reward reinforcement learning by approximating the domain models and the value function. In *International Conference on Machine Learning*, pp. 471–479.
- Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, 2(1):86–99.
- Tan, M. (1991). Learning a cost-sensitive internal representation for reinforcement learning. In L. A. Birnbaum and G. C. Collins (eds.), *Proceedings of the Eighth International Workshop on Machine Learning*, pp. 358–362. Morgan Kaufmann, San Mateo, CA.
- Tan, M. (1993). Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337. Morgan Kaufmann, San Mateo, CA.
- Taylor, G. and Parr, R. (2009). Kernelized value function approximation for reinforcement learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1017–1024. ACM.
- Taylor, M. E. and Stone, P. (2009). Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 10:1633–1685.
- Tesauro, G. J. (1986). Simple neural models of classical conditioning. *Biological Cybernetics*, 55:187–200. (cit. on p. 22).
- Tesauro, G. J. (1992). Practical issues in temporal difference learning. *Machine Learning*, 8:257–277. (cit. on p. 14).
- Tesauro, G. J. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219.
- Tesauro, G. J. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38:58–68. (cit. on p. 14).
- Tesauro, G. (2002). Programming backgammon using self-teaching neural nets. *Artificial Intelligence*, 134(1):181–199.
- Tesauro, G. J., Galperin, G. R. (1997). On-line policy improvement using Monte-Carlo search. In *Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference*, pp. 1068–1074. MIT Press, Cambridge, MA.

- Tesauro, G., Gondek, D. C., Lechner, J., Fan, J., and Prager, J. M. (2012). Simulation, learning, and optimization techniques in watson’s game strategies. *IBM Journal of Research and Development*, 56(3.4):16:1–16:11.
- Tesauro, G., Gondek, D. C., Lechner, J., Fan, J., and Prager, J. M. (2013). Analysis of WATSON’s strategies for playing Jeopardy! *Journal of Artificial Intelligence Research*, 21:205–251.
- Tham, C. K. (1994). *Modular On-Line Function Approximation for Scaling up Reinforcement Learning*. PhD thesis, Cambridge University.
- Thathachar, M. A. L. and Sastry, P. S. (1985). A new approach to the design of reinforcement schemes for learning automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 15:168–175.
- Thathachar, M. and Sastry, P. S. (2002). Varieties of learning automata: an overview. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(6):711–722.
- Thathachar, M. and Sastry, P. S. (2011). *Networks of learning automata: Techniques for online stochastic optimization*. Springer Science & Business Media.
- Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. (2015). Personalized ad recommendation for life-time value optimization guarantees. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-15)*.
- Thistlethwaite, D. (1951). A critical review of latent learning and related experiments. *Psychological Bulletin*, 48(2):97–129.
- Thomas, P. (2014). Bias in natural actor-critic algorithms. *International Conference on Machine Learning 31. JMLR W&CP* 32(1):441–448.
- Thomas, P. S. (2015). *Safe Reinforcement Learning*. PhD thesis, University of Massachusetts Amherst.
- Thomas, P. S., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 3000–3006. The AAAI Press, Palo Alto, CA.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294.
- Thompson, W. R. (1934). On the theory of apportionment. *American Journal of Mathematics*, 57:450–457.
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review, Series of Monograph Supplements*, II(4).
- Thorndike, E. L. (1911). *Animal Intelligence*. Hafner, Darien, CT. (cit. on p. 17).
- Thorp, E. O. (1966). *Beat the Dealer: A Winning Strategy for the Game of Twenty-One*. Random House, New York.
- Tian, T. (2017). Empirical Study of Sliding-Step Methods in Temporal Difference Learning. University of Alberta MSc thesis.
- Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop. COURSERA: Neural networks for machine learning.
- Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642–1645.
- Tolman, E. C. (1932). *Purposive Behavior in Animals and Men*. Century, New York.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208.

- Tsai, H.-S., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., and Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is surfacescient for behavioral conditioning. *Science*, 324(5930):1080–1084.
- Tsetlin, M. L. (1973). *Automaton Theory and Modeling of Biological Systems*. Academic Press, New York. (cit. on p. 20).
- Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202.
- Tsitsiklis, J. N. (2002). On the convergence of optimistic policy iteration. *Journal of Machine Learning Research*, 3:59–72.
- Tsitsiklis, J. N. and Van Roy, B. (1996). Feature-based methods for large scale dynamic programming. *Machine Learning*, 22:59–94.
- Tsitsiklis, J. N., Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690.
- Tsitsiklis, J. N., Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, 35:1799–1808.
- Turing, A. M. (1948). Intelligent Machinery, A Heretical Theory. *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*, 105. (cit. on p. 18).
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind* 433–460.
- Ungar, L. H. (1990). A bioreactor benchmark for adaptive network-based process control. In W. T. Miller, R. S. Sutton, and P. J. Werbos (eds.), *Neural Networks for Control*, pp. 387–402. MIT Press, Cambridge, MA.
- Urbanczik, R. and Senn, W. (2009). Reinforcement learning in populations of spiking neurons. *Nature neuroscience*, 12(3):250–252.
- Urbanowicz, R. J., Moore, J. H. (2009). Learning classifier systems: A complete introduction, review, and roadmap. *Journal of Artificial Evolution and Applications*. (cit. on p. 21).
- Valentin, V. V., Dickinson, A., and O’Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, 27(15):4019–4026.
- Van Roy, B., Bertsekas, D. P., Lee, Y., Tsitsiklis, J. N. (1997). A neuro-dynamic programming approach to retailer inventory management. In *Proceedings of the 36th IEEE Conference on Decision and Control*, Vol. 4, pp. 4052–4057.
- van Seijen, H. (2016). Ective multi-step temporal-difference learning for non-linear function approximation. arXiv preprint arXiv:1608.05151.
- van Seijen, H., Mahmood, A. R., Pilarski, P. M., Machado, M. C., and Sutton, R. S. (2016). True online temporal-difference learning. *Journal of Machine Learning Research* 17(145), 1–40.
- van Seijen, H. and Sutton, R. S. (2014). True online TD(λ). In *Proceedings of the 31st International Conference on Machine Learning*. JMLR W&CP 32(1):692–700.
- van Seijen, H., van Hasselt, H., Whiteson, S., Wiering, M. (2009). A theoretical and empirical analysis of Expected Sarsa. In *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, pp. 177–184.
- Varga, R. S. (1962). *Matrix Iterative Analysis*. Englewood Cliffs, NJ: Prentice-Hall.
- Vasilaki, E., Frémaux, N., Urbanczik, R., Senn, W., and Gerstner, W. (2009). Spike-based reinforcement learning in continuous state and action space: when policy gradient methods fail. *PLoS Computational Biology*, 5(12).
- Viswanathan, R. and Narendra, K. S. (1974). Games of stochastic automata. *IEEE Transactions on Systems, Man, and Cybernetics*, 4:131–135.

- Vlassis, N., Ghavamzadeh, M., Mannor, S., and Poupart, P. (2012). Bayesian reinforcement learning. In Wiering and van Otterlo (Eds.) *Reinforcement Learning: State-of-the Art*, pp. 359–386. Springer Berlin Heidelberg.
- von Neumann, J. and Morgenstern, O. (1944). *The Theory of Games and Economic Behavior*, Princeton University Press. (cit. on p. 24).
- Walter, W. G. (1950). An imitation of life. *Scientific American*, pages 42–45. (cit. on p. 18).
- Walter, W. G. (1951). A machine that learns. *Scientific American*, 185(2):60–63. (cit. on p. 18).
- Waltz, M. D., Fu, K. S. (1965). A heuristic approach to reinforcement learning control systems. *IEEE Transactions on Automatic Control*, 10:390–398. (cit. on p. 19).
- Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. Ph.D. thesis, Cambridge University. (cit. on p. 23).
- Watkins, C. J. C. H., Dayan, P. (1992). Q-learning. *Machine Learning*, 8:279–292.
- Werbos, P. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. Phd Thesis, Harvard University, Cambridge, Massachusetts.
- Werbos, P. J. (1977). Advanced forecasting methods for global crisis warning and models of intelligence. *General Systems Yearbook*, 22:25–38. (cit. on p. 23).
- Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick and F. Kozin (eds.), *System Modeling and Optimization*, pp. 762–770. Springer-Verlag, Berlin.
- Werbos, P. J. (1987). Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. *IEEE Transactions on Systems, Man, and Cybernetics*, 17:7–20. (cit. on p. 23).
- Werbos, P. J. (1988). Generalization of back propagation with applications to a recurrent gas market model. *Neural Networks*, 1:339–356.
- Werbos, P. J. (1989). Neural networks for control and system identification. In *Proceedings of the 28th Conference on Decision and Control*, pp. 260–265. IEEE Control Systems Society.
- Werbos, P. J. (1990). Consistency of HDP applied to a simple reinforcement learning problem. *Neural Networks*, 3:179–189.
- Werbos, P. J. (1992). Approximate dynamic programming for real-time control and neural modeling. In D. A. White and D. A. Sofge (eds.), *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, pp. 493–525. Van Nostrand Reinhold, New York.
- Werbos, P. J. (1994). *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting* (Vol. 1). John Wiley and Sons.
- White, A. (2015). *Developing a Predictive Approach to Knowledge*. Phd thesis, University of Alberta.
- White, A. and White, M. (2016). Investigating practical linear temporal difference learning. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, pp. 494–502.
- White, D. J. (1969). *Dynamic Programming*. Holden-Day, San Francisco.
- White, D. J. (1985). Real applications of Markov decision processes. *Interfaces*, 15:73–83. (cit. on p. 16).
- White, D. J. (1988). Further real applications of Markov decision processes. *Interfaces*, 18:55–61. (cit. on p. 16).

- White, D. J. (1993). A survey of applications of Markov decision processes. *Journal of the Operational Research Society*, 44:1073–1096. (cit. on p. 16).
- Whitehead, S. D., Ballard, D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, 7:45–83.
- Whitt, W. (1978). Approximations of dynamic programs I. *Mathematics of Operations Research*, 3:231–243.
- Whittle, P. (1982). *Optimization over Time*, vol. 1. Wiley, New York. (cit. on p. 17).
- Whittle, P. (1983). *Optimization over Time*, vol. 2. Wiley, New York. (cit. on p. 17).
- Wickens, J. and Kötter, R. (1995). Cellular models of reinforcement. In Houk, J. C., Davis, J. L., and Beiser, D. G., editors, *Models of Information Processing in the Basal Ganglia*, pages 187–214. MIT Press, Cambridge, MA.
- Widrow, B., Gupta, N. K., Maitra, S. (1973). Punish/reward: Learning with a critic in adaptive threshold systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 3:455–465. (cit. on p. 20).
- Widrow, B., Hoff, M. E. (1960). Adaptive switching circuits. In *1960 WESCON Convention Record Part IV*, pp. 96–104. Institute of Radio Engineers, New York. Reprinted in J. A. Anderson and E. Rosenfeld, *Neurocomputing: Foundations of Research*, pp. 126–134. MIT Press, Cambridge, MA, 1988. (cit. on pp. 18, 20).
- Widrow, B., Smith, F. W. (1964). Pattern-recognizing control systems. In J. T. Tou and R. H. Wilcox (eds.), *Computer and Information Sciences*, pp. 288–317. Spartan, Washington, DC. (cit. on p. 20).
- Widrow, B., Stearns, S. D. (1985). *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.
- Wiering, M., van Otterlo, M. (2012). *Reinforcement Learning*. Springer Berlin Heidelberg. (cit. on p. 23).
- Wiewiora, E. (2003). Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research* 19:205–208.
- Williams, R. J. (1986). Reinforcement learning in connectionist networks: A mathematical analysis. Technical Report ICS 8605. Institute for Cognitive Science, University of California at San Diego, La Jolla.
- Williams, R. J. (1987). Reinforcement-learning connectionist systems. Technical Report NU-CCS-87-3. College of Computer Science, Northeastern University, Boston.
- Williams, R. J. (1988). On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks*, pp. I263–I270. IEEE San Diego section and IEEE TAB Neural Network Committee.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.
- Williams, R. J., Baird, L. C. (1990). A mathematical analysis of actor-critic architectures for learning optimal controls through incremental dynamic programming. In *Proceedings of the Sixth Yale Workshop on Adaptive and Learning Systems*, pp. 96–101. Center for Systems Science, Dunham Laboratory, Yale University, New Haven.
- Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279.
- Wilson, S. W. (1994). ZCS: A zeroth order classifier system. *Evolutionary Computation*, 2:1–18.
- Wise, R. A. (2004). Dopamine, learning, and motivation. *Nature Reviews Neuroscience*, 5(6):1–12.

- Witten, I. H. (1976). The apparent conflict between estimation and control—A survey of the two-armed problem. *Journal of the Franklin Institute*, 301:161–189.
- Witten, I. H. (1977). An adaptive optimal controller for discrete-time Markov environments. *Information and Control*, 34:286–295. (cit. on p. 23).
- Witten, I. H., Corbin, M. J. (1973). Human operators and automatic adaptive controllers: A comparative study on a particular control task. *International Journal of Man-Machine Studies*, 5:75–104.
- Woodbury, T., Dunn, C., and Valasek, J. (2014). Autonomous soaring using reinforcement learning for trajectory generation. In *52nd Aerospace Sciences Meeting*, page 0990.
- Woodworth, R. S. (1938). *Experimental psychology*. New York: Henry Holt and Company. (cit. on p. 17).
- Xie, X. and Seung, H. S. (2004). Learning in neural networks by reinforcement of irregular spiking. *Physical Review E*, 69(4).
- Xu, X., Xie, T., Hu, D., and Lu, X. (2005). Kernel least-squares temporal difference learning. *International Journal of Information Technology* 11(9):54–63.
- Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., and Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science*, 345(6204):1616–1619.
- Yee, R. C., Saxena, S., Utgoff, P. E., Barto, A. G. (1990). Explaining temporal differences to create useful concepts for evaluating states. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 882–888. AAAI Press, Menlo Park, CA.
- Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476.
- Young, P. (1984). *Recursive Estimation and Time-Series Analysis*. Springer-Verlag, Berlin.
- Yu, H. (2010). Convergence of least squares temporal difference methods under general conditions. *International Conference on Machine Learning* 27, pp. 1207–1214.
- Yu, H. (2012). Least squares temporal difference methods: An analysis under general conditions. *SIAM Journal on Control and Optimization*, 50(6), 3310–3343.
- Yu, H. (2015a). On convergence of emphatic temporal-difference learning. ArXiv:1506.02582. A shorter version appeared in *Conference on Learning Theory 18, JMLR W&CP* 40.
- Yu, H. (2015b). Weak convergence properties of constrained emphatic temporal-difference learning with constant and slowly diminishing stepsize. ArXiv:1511.07471.
- Zhang, M., Yum, T. P. (1989). Comparisons of channel-assignment strategies in cellular mobile telephone systems. *IEEE Transactions on Vehicular Technology*, 38:211–215.
- Zhang, W. (1996). *Reinforcement Learning for Job-shop Scheduling*. Ph.D. thesis, Oregon State University. Technical Report CS-96-30-1.
- Zhang, W., Dietterich, T. G. (1995). A reinforcement learning approach to job-shop scheduling. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pp. 1114–1120. Morgan Kaufmann.
- Zhang, W., Dietterich, T. G. (1996). High-performance job-shop scheduling with a time-delay TD(λ) network. In D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (eds.), *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1024–1030. MIT Press, Cambridge, MA.
- Zweben, M., Daun, B., Deale, M. (1994). Scheduling and rescheduling with iterative repair. In M. Zweben and M. S. Fox (eds.), *Intelligent Scheduling*, pp. 241–255. Morgan Kaufmann, San Francisco.