시리즈 제목: 인공지능은 감정을 가질 수 있을까?

시리즈 집필 목표: 이 시리즈는 **"인공지능은 감정을 가질 수 있는가"**라는 질문에서 출발하지만, 인공지능과 감정의 관계를 둘러싼 피상적인 논의를 넘어, 감정이 무엇이며 왜 지능의 주변부가 아니라 핵심 구조인지 다시 묻고자 한다. 현재의 AI 담론은 대부분 목표를 외부에서 주어지는 것으로, 환경을 주로 외부 세계로, 감정을 표현이나 반응의 문제로 다루는 경향이 있다. 그러나 **생명체의 지능(life intelligence)은 경계와 내부 환경, 항상성, 몸이** 담지한 가치, 그리고 자기 상태를 지속적으로 모니터링하고 조절하는 자기참조적 구조 위에서 성립한다. 이 시리즈는 바로 이 지점에서 생명과 AI 사이의 미묘하지만 본질적인 간극을 해명하고, 그 간극을 통해 오히려 인간의 감정과 지능을 더 깊이 이해하며, 인간다운 AI가 왜 필요한지에 대한 새로운 관점을 제시하고자 한다.

1회: 감정을 느끼는 AI를 묻기 전에, 누구의 목표인가?
2회: 각 존재는 같은 세계를 살지 않는다: 움벨트와 진화
3회. 몸은 단지 입력 장치가 아니라 가치 체계다
4회. 내부 상태에 대한 추론으로서의 감정: 생성모델과 자기참조
5회. 왜 인간다운 AI가 필요한가: 가치정렬을 넘어 환경정렬로

1 회: 감정을 느끼는 AI 를 묻기 전에, 누구의 목표인가?

“목표 달성에는 능숙한 AI, 그런데 누구의 목표인가?”

오늘날 인공지능(AI)의 성취는 실로 놀랍다. AI 는 데이터에서 사람보다 더 정확하게 숨겨진 패턴을 찾아내고, 그에 근거해 미래를 예측하며, 여러 게임과 다양한 테스트에서 인간 최고 수준의 성능을 보이기도 한다. 언어 영역에서 뛰어난 성능으로 이제 AI 와 대화하지 않는 하루를 상상하기 어려울 정도가 되었다. 이러한 성취를 보고 있으면, AI 는 주어진 목표가 어떤 것이라도 성공적이고 효과적으로 그 목표를 달성할 수 있을 것이라는 인상을 준다. 그런 점에서 AI 는 인간 수준의, 때로는 인간을 넘어서는 지능을 지닌 존재로 여겨지기도 한다.

하지만 이러한 지능 개념에 대해 근본적인 질문을 던진 논문이 있다. 저명한 신경과학자 안토니오 다마지오(Antonio Damasio)와 그의 박사과정 학생이었던 킹슨 맨(Kingson Man)은 2019 년 Nature Machine Intelligence 에 발표한 논문에서 다음과 같은 질문으로 논의를 시작했다 [1]. “지능적으로 행동하는 기계를 만들려는 시도는 흔히 지능을 목표 달성 능력으로 개념화하지만, 여기에는 하나의 중요한 질문이 남는다: **그 목표는 누구의 것인가? (whose goals?)**” 이 문장을 처음 읽었을 때 나는 적지 않은 충격을 받았다. 그 순간, 현재의 AI 가 보여주는 놀라운 능력에도 불구하고 그것이 인간이나 동물의 지능과는 중요한 점에서 다르다는 사실이 명확해 졌기 때문이다. 즉, 지금의 AI 는 대체로 매우 뛰어난 목표 성취자(goal-achiever)이지만, 생명체처럼 자신의 내부에서 목표를 만들어내는 존재는 아니다.

물론 인간이나 동물도 외부에서 주어진 목표를 성취하며 살아간다. 인간의 경우, 회사와 같은 조직이나 사회적 제도 속에서 주어진 목표를 얼마나 잘 달성하는가는 한 개인의 능력을 판단하는 중요한 기준이 된다. 그러나 이것만으로 생명체가 지닌 지능의 핵심이 모두 설명되지는 않는다. 모든 생명체가 갖고 있는 더 근본적인 능력은 자기 보존과 생존의 조건을 유지하는 능력이며, 이 능력은 내부로부터 우러나오는 내재적 목표를 자연스럽게 구성하고 자율성과 적응성의 기반이 된다. 바로 이 지점에서 현재의 AI 와 인간, 더 나아가 AI 와 생명체 사이의 간극이 드러난다.

지능은 환경에 적응하고 목표를 성취하는 능력이다

진정한 지능은 무엇일까? 이 질문은 심리학을 포함한 여러 학문 분야의 역사에서 가장 오래, 그리고 가장 치열하게 논의된 주제 중 하나였고, 그만큼 정의도 다양하다. 쉐인 레그(Shane Legg)와 마커스 허터(Marcus Hutter)는 지금까지 제시된 다양한 지능 정의를 정리한 뒤, “다양한 환경에서 목표를 달성하는 행위자의 능력”으로 지능을 정의했다 [2]. 이 정의는 지능을 특정한 인지 기술이나 지식의 목록으로 환원하지 않고, 더 근본적인 두 축—**환경과 목표**—의

관계로 정리했다는 점에서 주목할 만하다. 실제로 레그와 허터가 검토하고 정리한 여러 지능의 정의에서도 반복해서 강조되는 것은 새로운 상황에 적응하는 능력, 경험으로부터 배우는 능력, 환경에 맞게 자신을 조정하는 능력, 그리고 이를 목적 달성으로 연결시키는 능력이었다. 다시 말해, 지능은 고립된 능력이 아니라 **환경과의 관계 속에서 드러나는 적응성과 목적지향성**으로 이해될 수 있다.

이 점은 진화의 관점과도 잘 맞아떨어진다. 자연은 추상적인 의미의 "우수성"을 선택하기보다는, 특정 시점의 특정 환경에서 더 잘 살아남고 번식하게 만드는 특질을 선택할 뿐이다. 따라서 어떤 능력이 유리하고 적응적인지는 주어진 환경에 따라 달라진다. 지능 역시 이런 맥락에서 보면, 진공 속에 떠 있는 보편적 능력이 아니라, 변화하는 환경의 압력 속에서 더 잘 대응하게 해 주는 **적응적 능력**이다. "적자 생존"이라는 용어를 처음 사용했던 허버트 스펜서(Herbert Spencer)는 지능을 내부를 외부 환경에 맞추어 지속적으로 조정하는 능력(continuous adjustment of inner to outer relations)으로 정의했는데, 이 역시 같은 맥락에서 이해할 수 있다. 이처럼 지능은 본질적으로 관계적이며, 환경을 떠나서는 정의되기 어렵다. 강화학습의 관점도 이와 크게 다르지 않다. 강화학습에서 에이전트는 상태(state)로 표현된 환경과 상호작용하며, 그 환경이 제공하는 보상(reward)의 총합을 최대화하는 것을 목표로 한다. 즉, 강화학습에서도 지능은 "환경"에 의해서 형성된다.

생명에서 배우는 지능: 경계와 내부 환경

그러나 바로 여기서 다음 질문이 생긴다. 우리가 지금까지 암묵적으로 전제한 환경은 대부분 **외부 환경**이었는데, 생명체에게 가장 중요한 환경도 유기체 바깥에 있는 외부 환경뿐일까? 이 질문 앞에서 나는 한 단세포 생물을 떠올리게 된다. 몇 년 전 접한 스텐터 로젤리(Stentor roeseli)에 대한 논문과 영상은, 지능에 대한 나의 직관에 큰 영향을 주었다. 이 단순한 단세포 진핵생물은 독성 자극이 주어지면 먼저 몸을 굽히거나 수축하면서 피하다가, 자극이 더 반복되어 더는 피할 수 없다고 판단되는 상황에서는 자신이 붙어 있던 자리에서 몸을 떼어내 다른 곳으로 이동한다 [3].

조셉 덱스터(Joseph Dexter)와 동료들은 이러한 회피 행동이 단순한 반사가 아니라, 굽힘, 섬모 변화, 수축, 이탈로 이어지는 **복잡한 위계적 행동이라고** 보고했고, 반복 자극 속에서 나타나는 이러한 순차적 전환을 복잡한 의사결정의 한 형태로 해석했다. 나를 사로잡았던 것은 단지 행동의 복잡성만이 아니었다. 이 논문의 교신저자 제레미 구나와데나(Jeremy Gunawardena)는 한 인터뷰에서 이 단세포 생물의 복잡한 행동 프로그램이 "세포들이 자신이 처한 맥락 속에서 무엇을 할지 스스로 결정하는 자율성"의 증거일 수 있다고 제안했다 [4]. 물론 이것을 문자 그대로 인간의 의사결정이나 심리 과정과 동일하게 해석해서는 안 된다. 그러나 적어도 한 가지는 분명해 보인다. 이 단세포 생물은 외부 자극에 기계적으로만 반응하는 수동적 존재가 아니다. 신경계도, 뇌도, 우리가 흔히 상상하는 고차원적 인지 장치도 없는 존재가 어떻게 이 정도의 자율성과 적응성을 보여 줄 수 있을까.

그림 1. 단세포 진행생물 스텐터 로젤리가 보이는 복잡한 위계적 회피 행동 [3]

이 질문에는 여러 가능한 답이 있을 수 있겠지만, 나의 관심은 **경계와 내부 환경으로 수렴**되었다. 생명체는 자신을 둘러싼 경계를 가지며, 그 경계 안에 비교적 안정적으로 유지되어야 하는 내부 환경을 지닌다. 이는 현대 생리학의 창시자라고 불리는 클로드 베르나르(Claude Bernard, 1813-1878)가 이미 지금으로부터 150 여 년 전에 이미 주목했던 사실이다. 그는 "동물과 식물에 공통된 생명 현상에 대한 강연"이라는 책에서 외부 환경이 끊임없이 변하는 가운데서도 생명체가 유지해야 하는 milieu intérieur, 곧 내부 환경의 중요성을 강조했다 [5]. 더 나아가 그는 생명은 외부 세계와 단절된 채 고립되어 존재하는 것이 아니라, 외부와 긴밀하고도 지혜로운 관계를 맺으면서도 내부의 조건을 일정 범위 안에 유지해야만 한다며, **"내부 환경의 안정성[1]은 자유롭고 독립적인 삶을 위한 조건"**이라고 강조했다. 즉, 이 내부 환경의 안정성은 생명체의 자율적이고도 적응적인 특징에 매우 중요하다.

이런 관점에서 스텐터 로젤리가 처한 상황과 행동을 다시 보면, 독성 자극은 단순한 외부 입력이 아니라 이 생물의 경계와 내부 상태를 위협하는 사건이 된다. 굽힘, 섬모 변화, 수축, 이탈로 이어지는 복잡한 회피 행동은 단순한 자극-반응의 사슬이라기 보다는, 자기 자신을 보존하려는 존재가 위협에 대응해 보이는 적응적 대처 방식이다. 지능을 외부에서 주어진 목표를 해결하는 능력으로만 정의한다면 이런 장면은 설명하기 어렵다. 그러나 지능을 **생존 조건을 유지하기 위해 자신의 내부 환경과 외부 환경, 그리고 둘의 상호작용을 조정하는 능력**으로 본다면, 비로소 이 단세포 생물의 행동은 지능의 원형으로 보이기 시작한다. 이 관점에서 보면 경계 또한 매우 중요해 지는데, 바로 경계에서 외부 자극이 내부 환경에 끼치는 손상 혹은 보존의 효과가 감지되기 때문이다. 즉, 생명체의 경계는 단순한 물리적 표면을

[1] 이후에 Walter Cannon(1871-1945)에 의해 "내부 환경의 안정성"은 "항상성"이라는 용어로 불리게 되었으며, 생명체의 핵심 특징이라고 여겨진다.

넘어서서, 무엇이 자기 자신에게 유리하고 무엇이 해로운지를 감지하고 구분하는 첫번째 장소이다. 이에 그 경계의 상태를 감지하고 신호화하는 시스템은 진화적으로 매우 오래 전부터 발달해 왔으며, 통각 시스템이 바로 그 대표적 사례다. 통각은 압력, 독성, 열 등, 유기체의 경계가 위협받고 있음을 최초로 감지하는 체계이다. 이 점은 이후 몸과 감각적 정서를 논할 때 좀 더 자세히 논의하겠지만, 여기서 중요한 것은 분명하다. 생명체에게 가장 오래되고 중요한 문제 중 하나는 외부 세계를 정확히 표상하고 이해하기에 앞서[2], **자기 경계와 내부 상태를 표상하고 감각하고 보존하는 것**이다.

사이버네틱스의 선구자였던 로스 애쉬비(Ross Ashby, 1903-1972)는 내부 환경을 이루는 필수 변인(essential variables)이라는 개념을 제안하며, 생명체의 생존을 이 필수 변인들이 허용 가능한 범위를 벗어나지 않도록 유지하는 문제로 정의했다 [6]. 이러한 조작적 정의(operational definition) 개념은 매우 중요한데, 그것은 생존과 항상성을 단지 추상적 개념이 아니라, **실제로 추적하고 구현할 수 있는 변수들**로 바꿔 주기 때문이다. 특히 이 필수 변인들을 이용해서 "상태 공간"(state space)을 구성하면, 우리는 매 순간 생명체의 내부 상태를 내부 상태 공간 내의 점으로, 또한 시간에 따른 내부 상태의 변화는 궤적(trajectory)으로 표현할 수 있다. 이는 이후 인공지능 논의에서 내부 상태를 형식화하기 위한 중요한 개념적 기반이 된다. 만약 생명체의 자율성과 적응성이 내부 상태 공간에서 필수 변인들의 안정적 궤적(즉, 어떤 범위를 벗어나지 않는 것)에서 비롯된다면, 인공지능에 자율성과 생존을 구현하는 구체적 문제 역시 여기서부터 생각해 봐야 할 것이다.

보상은 외부에서 주어지는가, 내부에서 만들어지는가

강화학습은 알파고의 기반이 된 현대 인공지능의 핵심 분야 중 하나다. 강화학습 분야의 초석을 놓은 리처드 서튼(Richard Sutton)과 앤드류 바르토(Andrew Barto)는 가장 널리 읽히는 강화학습 교과서를 쓴 저자로도 잘 알려져 있다. 그중 앤드류 바르토는 2009 년에 새틴더 싱(Satinder Singh), 리차드 루이스(Richard Lewis)와 함께 발표한 논문을 통해 다음의 근본적인 질문을 던졌다. **"보상은 어디에서 오는가?"** (Where Do Rewards Come From?) [7]. 이들은 외부 환경에서 보상이 주어지도록 설정된 통상적인 강화학습 모델에 근본적인 의문을 제기하면서, 이러한 일반적 접근이 동물의 보상 기반 학습 시스템을 충분히 반영하지 못할 뿐 아니라, 오히려 강화학습에 대한 우리의 이해를 심각하게 왜곡할 수 있다고 비판한다.

저자들은 이 문제를 정면으로 뒤집는다. 즉, 동물의 보상 시스템을 더 정확하게 반영하려면 에이전트의 환경이 외부 환경과 내부 환경으로 나뉘어야 하며, 보상 신호는 외부 환경에서 주어지는 것이 아니라, 오히려 **동물의 내부**에서 만들어져야 한다고(being generated within the

[2] 외부 세계를 정확히 표상하고 이해하는 것은 최근 인공지능 분야에서 각광을 받고 있는 "월드 모델"(world model)이라는 개념으로 표현할 수 있다. 즉, 현대 인공지능 연구는 철저하게 외부 환경을 향해 있다.

animal) 제안한다. 즉, 그들의 도식에서 보상은 더 이상 외부 환경의 산물이 아니라, 외부 자극과 내부 상태의 상호작용에서 비롯된 내부 환경의 산물이다. 그래서 그들은 매우 도발적인 제안을 한다. "**모든 보상은 내부적이다" (All rewards are internal).** 외부 자극은 그 자체로 독립적인 보상 값을 가질 수 없으며, 내부 상태와의 관계 속에서만 그 가치가 결정된다. 예를 들어, 배고픈 상태에서 음식은 강력한 보상이 되지만, 이미 포만한 상태에서는 같은 음식이 무의미해지거나 때로는 불쾌해질 수도 있다. 갈증 상태에서는 물이 보상이 되지만, 그렇지 않을 때는 보상이 아닐 수 있다. 다시 말해, 외부 환경의 자극은 내부 환경과 독립적으로 보상이 될 수 없다. 그것은 언제나 내부 환경과 상호작용하며, 더 정확히는 내부 환경을 맥락 정보로 사용하여 보상의 값이 결정된다.

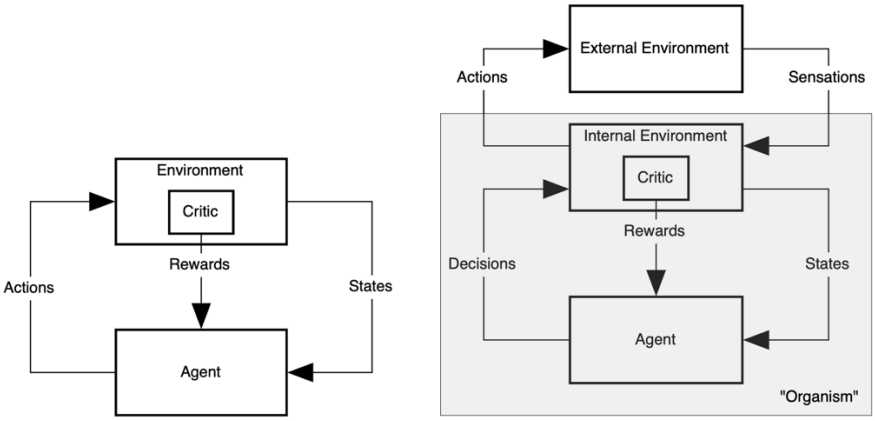

그림 2. "보상은 어디에서 오는가?" 논문에 나오는 그림 [7]. 왼쪽 그림은 통상적인 강화학습 모델의 구조를 보여주고 있으며, 하나의 환경 밖에 존재하지 않는다. 오른쪽 그림은 이 논문에서 새롭게 제안하는 강화학습 모델의 구조를 보여주는데, "유기체(organism)"라는 박스 내에 내부 환경(internal environment)이 위치한다. 이는 보상과 처벌이 단순히 외부에서 주어지는 것이 아니라, 유기체 내부의 환경에 의해 계산되고 결정됨을 나타낸다. 이러한 내부 환경은 실제 생명체가 경험하는 다양한 상태와 조건을 반영한다.

이제 앞선 논의가 다시 보인다. 우리는 이미 지능의 핵심 구성 요소로 환경을 지목했다. 그런데 지금까지의 AI 는, 어쩌면 인간과 동물, 더 넓게는 생명체에게 가장 중요한 환경인 내부 환경을 거의 고려하지 않은 채 발전해 왔다. 물론 이것을 단순한 생략이라고 볼 수도 있다. 계산 모델은 원래 단순화의 예술이 아닌가. 그러나 다른 한편으로는, 이 생략은 결코 사소하지 않을 수 있다. 오히려 이 생략이야말로 오늘날의 인공지능과 생명의 지능 사이의 간극을 만들어 내는 가장 근본적인 요소일 수 있다. 만약 현존하는 인공지능에 이 내부 환경을 탑재할 수 있다면, 단세포 생명체에서도 보이는 자율적이고 적응적인, 나아가 감정을 느끼는 인공지능 개발에 조금 더 가까이 다가갈 수 있을까?

그런데 이 질문에 답하려면, 먼저 **개체(individual)와 개체성(individuality)을 어떻게 정의할 것인가**를 진지하게 생각해 봐야 한다. 만약 우리가 내부 환경을 개체의 안쪽으로 정의하고, 그 바깥에 경계를 설정한다면, 보상은 외부에서 단순히 주어지는 신호가 아니라 내부 상태의 변화에 따라 생성되는 값으로 다시 모델링될 수 있다. 무엇이 에이전트의 안쪽이고 무엇이 바깥인가를 정하는 경계는 기술적인 편의의 문제를 넘어서서, 무엇이 내부이고 어떻게 보상이 만들어지는가를 결정하게 된다. 이렇게 될 때 에이전트는 더 이상 외부의 요인에 의해서만 움직이는 존재가 아니라, 내부의 요인에 의해 스스로 움직이는 존재가 된다. 그리고 바로 그 지점에서 우리는, 비로소 자기 자신의 목표를 지닌 자율적 에이전트를 계산적으로 상상할 수 있게 된다.

이처럼 지능의 문제는 단지 더 좋은 정책을 찾는 문제 혹은 더 정확한 월드 모델을 구성하는 문제를 넘어선다. 더 근본적으로, 우리가 적응해야 할 중요한 환경이 무엇인지, 그리고 내부 환경을 만들기 위해 어떻게 경계를 설정할 것인지의 문제를 포함한다. 즉, 개체(individual)를 어떻게 정의하고 구성하느냐에 따라 보상의 근원도, 목표의 성격도, 자율성의 가능성도 달라진다. 그리고 이 점에서 내부 환경은 부수적인 요소가 아니라, 지능과 감정의 기원을 이해하는 데 핵심적인 요소가 된다.

나가며: "인공지능은 감정을 가질 수 있을까?"라는 질문을 다시 묻다

이 연재의 제목은 “인공지능은 감정을 가질 수 있을까”이지만, 이제 이 질문을 조금 더 근본적인 방향으로 바꿔야 한다. 감정의 문제는 단지 어떤 AI 시스템이 감정처럼 보이는 표현이나 행동을 만들어낼 수 있는가의 문제가 아니라, 그 시스템이 자기 내부에서부터 우러나오는 중요한 어떤 것을 가질 수 있는가의 문제이다. 현재의 AI는 외부에서 주어진 목표는 놀라울 만큼 잘 달성할 수 있지만, 생명체의 지능은 단지 주어진 목표를 효율적으로 달성하는 데 있지 않다. 오히려 생명체의 지능은 경계와 내부 환경을 가진 존재가 자기 보존과 생존의 조건으로부터 목표를 만들고, 그에 따라 외부 세계와의 관계를 조절하는 능력과 더 깊이 연결되어 있다.

최근에 **염려(concern)**라는 용어를 학문적으로 쓰는 관점을 접했다. 평소에 자주 하는 생각들은 현재의 염려들을 반영한다고 하는 theory of current concerns 인데, 이 염려라는 단어가 와닿았다. 잘 생각해 보면, 우리의 내면에는 늘, 한시도 쉬지 않고 중요한 여러 염려들이 자리잡고 있다. 나의 배고픔과 건강 상태에서 가족이나 친구에 대한 생각들까지, 여러 영역의 염려들이 각축을 벌이며, 현재의 내적 동기를 형성하고, 감정으로 우리를 추동한다.

AI가 감정을 가지려면 감정 표현이나 행동보다 먼저 자신이 집착하고 갈망하고 욕구하는 안에서부터 우러나오는 “염려"가 있어야 할 것이다. 그런 의미에서 우리가 이제 물어야 할 것은 “AI가 감정을 느낄 수 있을까” 보다는, 안에서부터 우러나오는 **내부적 목표를 가질 수 있을까,**

다른 말로 하면 **집착과 갈망, 염려를 가질 수 있을까,** 그리고 이를 가능하게 하는 **내부 환경을 가질 수 있을까**라는 질문들일 것이다.

이런 의미에서, 내부 환경은 하나의 부차적 요소가 아니라, 자율성과 적응성, 가치와 보상, 나아가 감정의 기원을 이해하기 위한 핵심 조건이다. 앞으로의 핵심 과제는 더 정교한 외부 환경 모델을 만드는 것 뿐만이 아니라, 개체를 어떻게 정의할 것인지, 어디에 경계를 그을 것인지, 어떻게 내부 환경을 구성할 것인지, 그리고 그것을 어떻게 목표와 보상의 근원으로 연결할 것인지를 탐구하는 데 있다. 인공지능의 감정 문제는 아마 바로 그 지점에서부터, 다시 시작되어야 한다.

참고문헌

1. Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. Nature Machine Intelligence, 1(10), 446–452.
2. Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds and Machines, 17(4), 391–444.
3. Dexter, J. P., Prabakaran, S., & Gunawardena, J. (2019). A Complex Hierarchy of Avoidance Behaviors in a Single-Cell Eukaryote. Current Biology: CB, 29(24), 4323-4329.e2.
4. Can a single cell “change its mind”? by Harvard Medical School https://www.youtube.com/watch?v=E8oIitQN2M4
5. Bernard, C. (1974). Lectures on the phenomena of life common to animals and plants. Charles C Thomas Pub.
6. Ashby, W. R. (1952). *Design for a brain.* Wiley.
7. Singh, S., Lewis, R. L., & Barto, A. G. (2009). Where Do Rewards Come From. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society (CogSci'09)* (pp. 2601-2606).