어떻게 인공지능에게 감정을 부여할 것인가?

인공지능에게 ‘온전한 감정’을 부여: 강인공지능 – 인공 의식의 영역

**의식이란**

the state of being awake and aware of one's surroundings – 주변을 인지할 수 있는 “깨어있음”의 성질을 의식이라 일컬음

의식과 무의식?

\*인공 의식의 경우 최첨단 신경과학, 컴퓨터공학으로 해결하지 못하는 문제(뇌의 구조와 작동 원리가 온전히 밝혀지지 않은 상태에서 해당 분야의 연구는 굉장히 어려울 수밖에 없다)

**감정이 어떻게 생겨나는가: 제임스-랑게 이론**

William James: 인간의 감정은 인간의 신체적 변화로부터 발현된다. 일례로, 떨리는 입술, 후들거리는 다리, 가빠지는 숨결과 고동치는 심박 등이 없다면 ‘두려움’이란 감정에는 무엇이 남는가? 또한 감정에 특정한 종류가 있어서 특정 신체적 반응이 특정 감정의 분야(기쁨, 슬픔 등과 같은)로 직결되는 것이 아닌, 특정 신체적 반응이 그때그때의 감정을 불러일으킨다고 기술했다…고한다.

개인적 의견: 신체적 변화로부터 감정이 발현되는 것이 아니라 감정의 변화에 의해 신체변화가 발현되는 것은 아닐까? – 이 ‘개인적 의견’은 이후 ‘일반적인 생각’으로 제시되고, 제임스는 자신의 이론이 일반적인 생각과는 대비된다고 설명. 아..네…

반론: 사람은 감정이 있어도 그것을 참을 수 있으며, 2009년의 한 연구의 결과 (뇌의 손상으로) 감정결핍이 있어도 신체는 반응하며, 역시 행동하지 않아도 감정이 생긴다는 결론이 있었다.

이론의 결론?: 감정은 생리현상 + α 로부터 생겨난다는 것이 학계의 두루미. 생리현상은 감정표현의 강도를 변화시킨다. 예) 슬퍼서 울기 시작했더니 더 슬퍼짐

생각해보자: 감정이 정말 신체의 생리현상으로부터 생기겠는가?

\*제임스-랑게 이론: 화날만한 일이 생겼는데 일단 대수롭지 않게 넘김-몸이화남-생각해보니빡치네?

\*현실에선 어떤가: 화날만한 일이 생겼는데 일단 대수롭지 않게 넘김-생각해보니빡치네?-몸이화남

\*이불을 차고 화가 나는가 화가 나서 이불을 차는가?

더 조사한 결과 제임스-랑게 이론은 감정이 왜 생기는가에 대한 여러가지 이론의 줄기 중 인체의 현상이 감정의 유발에 지대한 영향을 준다는 somatic theory(굳이 해석하자면 인체 이론쯤 정확한 해석은 모루겟고 ㅎㅎ)중의 하나이다. 내 의견에 가장 가까운 것은 **cognitive theory(인지 이론)**으로, 외부 환경에 대한 인지와 자기 인식의 판단이 감정을 유발하고, 신체적 반응을 이끌어낸다는 이론이다. 예를 들어 1. 28동 건물을 본다. 2. 뇌가 그 추한 몰골을 보고 역겹다고 판단하여 얼굴에 명령을 내린다. 3. 얼굴을 찌푸리며 지나간다/얼굴을 몹시 찌푸리며 수업을 들으러 들어간다.

**감정은 의식으로부터 비롯되는가?**

자기 인식(self-awareness): 자기 자신이 의식을 가짐을 인지하는 능력.

감정을 유발하는 것은 자기 인식의 영역 - 의식보다 한 단계 위.

**본론: 그래서 감정이 어떻게 생기는지를 무슨 수로 실험할 것이고 이를 어떻게 인공 지능에게 이식할 생각이라도 해 볼 생각이냐?**

두가지 한계점

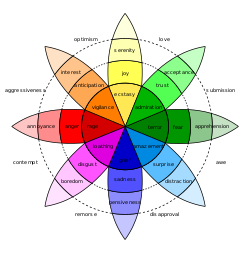
1. 현재 인간의 자기 인식은커녕 의식도 어떻게 돌아가는지 완전하게 파악하지 못한 상황에서 감정이 어떻게 생기는지 일개 학부 신입생이 알아내기란 불가능(세상에 불가능은 없다니까 0.0000000000001프로가능?)

2. 자기 인식이 어떻게 돌아가는지를 알더라도 이를 컴퓨터상으로 구현하기는 매우 어려울 것

**그래서 어떻게 할 거냐?**

튜링 테스트: 앨런 튜링이 고안한 컴퓨터의 지능 시험 방법. 질의자 하나가 둘에게 질문을 하는데 하나는 컴퓨터고 하나는 사람이다. 질의자는 뭐가 컴퓨터인지 모름. 질의자가 질문해보면서 뭐가 컴퓨터인지 구별해낼 수 없으면 테스트 통과. 의의는 **‘인간에 대한 본질을 깨우치기는 이르니 일단 겉보기에 사람 같으면 지능이 있다고 간주하면 어떠한가.’**

인식의 전환: 사람의 감정이 어떻게 생기는지도 정확히 모르고, 어떻게 생기는지 알더라도 이를 컴퓨터에 옮기기엔 영역이 너무 달라 힘들다면, 감정이 생기는 방식을 단순화하여 컴퓨터처럼 입출력의 개념으로 이해하면 어떨까?

로버트 플럿츀의 ‘감정의 바퀴’

인간의 기본적인 감정을 ‘믿음, 기쁨, 기대감, 분노, 슬픔, 역거움, 놀라움, 두려움’의 8가지로 구분하여 설명. 이 중 가장 기본이 **‘기쁨, 슬픔, 분노, 두려움’**의 4가지. 실험을 단순화하기 위해 가장 기본적인 4가지 감정만을 바탕으로, 인지 이론을 바탕으로 한 사고실험/경험 기록 진행.

실험 예시

00한 상황은 ∆∆에 의해 ㅁㅁ한 감정을 불러일으킨다.

기록 예시

00때문에 ∆∆하다고 생각하여 ㅁㅁ한 감정을 느꼈다.

예) 옆방분이 더럽게 부르지도 못하는 노래를 목청껏 불렀기 때문에, 내 고막이 손상되고 숙면 또는 공부/과제 진행에 방해가 된다고 생각하여, ‘분노’를 느꼈다.

아주 간단한 방식이지 아니한가. 컴퓨터로 치면 00은 입력이고, ∆∆는 입력이 주체에게 미치는 영향으로 실질적으로 알고리즘을 갖는 함수에 입력되는 부분일 것이다. 이것이 실현되려면 컴퓨터는 입력이 자신에게 어떠한 영향을 미쳤는지 파악할 수 있어야 한다(즉, 단순한 인지 능력이 요구된다.) ㅁㅁ는 입력에 의해 생긴 감정의 출력이다. ∆∆로부터 ㅁㅁ이 도출도는 가정이 바로 컴퓨터의 알고리즘이 될 것이다. 실험/기록의 종합에서의 중요한 것은 위의 ∆∆를 체계적으로 정리해 보는 것이다. 어떤 영향이 어떤 감정을 불러일으키는지 보자. 결과는 아래처럼 정리될 것이다.

|  |
| --- |
| 기쁨 |
| 00때문에 ∆∆하다고 생각하여 기뻤다. (여러가지 예시) |
| 슬픔 |
| 00때문에 ∆∆하다고 생각하여 슬펐다. (너도 여러가지) |
| 두려움 |
| 00때문에 ∆∆하다고 생각하여 두려웠다. (너두?) |
| 분노 |
| 00때문에 ∆∆하다고 생각하여 분노했다. (야나두) |

실험/기록의 결과를 충분히 얻었다고 생각할 때 각 감정의 항목들 내의 ∆∆를 추려 공통적인 부분들을 발견하고 정리한다. 출력은 정해져 있으므로 ∆∆들을 출력에 연결하면? 상당히 정형화된 블랙 박스 모델의 튜링 기계(어떤 종류의 입력이든 출력이든 넣어볼 수 있는 참 편한 기계)에 넣어볼 수 있을 것이다! 와아! 예상되는 예시를 하나 보자.

어떠한 놀라운 계산을 수행하는 튜링 기계

옆방분이 괴상망측한 노래를 한다

분노!

너무 속 편한 결과라 하면 할말이 없지만 뭐… 음. 최종 목표는 튜링 기계가 어떠한 식의 연산을 해야 할 지 개요를 짜는 것이다. 아직 안 했다. 할 예정이다. 오늘은 여기까지