인공지능은 사람일…까?



김현성

*주의. 본 글에는*

*영화 <2001: 스페이스 오디세이>(1968, 스탠리 큐브릭),*

*영화 <월-E>(2008, 앤드류 스탠튼),*

*게임 <Portal>(2007, Valve),*

*게임 <Portal 2>(2011, Valve)*

*의 스포일러가 포함되어 있습니다.*

우선 위에 써 있는 것처럼 이 글에는 몇몇 작품의 스포일러가 포함되어 있으니 주의하길 바란다. 혹시나 저 작품들 중 하나라도 감상하거나 플레이 해볼 생각이 조금이라도 있다면 이 다음 글로 넘기길 추천한다.그냥 스포일러 정도가 아니라 조금이라도 알면 정말 작품이 재미도 없고 감동도 없어지기 때문이다. 그걸 내가 어떻게 아냐고? 나도 알고 싶지 않았다…….

또한 주제가 주제이니만큼 이 글에는 인공지능과 관련하여 몇 가지 용어가 등장한다. 그 중 제일 중요한 용어인 **강인공지능**과 **약인공지능**에 대해서만 미리 설명하고 본격적으로 시작해보겠다. 그다지 어렵지는 않으니 도망가지는 말아주시길. 인공지능은 다음과 같이 둘로 분류할 수 있다. 먼저 강인공지능은 실제 인간처럼 특정되지 않은 모든 일을 수행할 수 있는 인공지능을 말한다. 터미네이터나 자비스 같은 대부분의 영화 속 인공지능이 여기에 해당되며, 앞으로 글에서 언급될 인공지능들은 대부분이 강인공지능에 속한다. 반대로 약인공지능은 특정 한 가지의 일만 수행할 수 있는 인공지능으로, 제일 유명한 예시로는 알파고나 이미지 인식 인공지능 같은 친구들이 있다.

강인공지능이 기술적으로 가능한지는 여전히 의견이 분분하다. 하지만 내가 이 글에서 하고 싶은 이야기는 강인공지능이 실재할 때 벌어질 일에 대한 것들이므로 강인공지능이 실현 가능하다는 가정을 하고 시작하겠다.

— ◇ —

빨간색, 옅은, 신호등, 어두운이(가) 표시된 사진

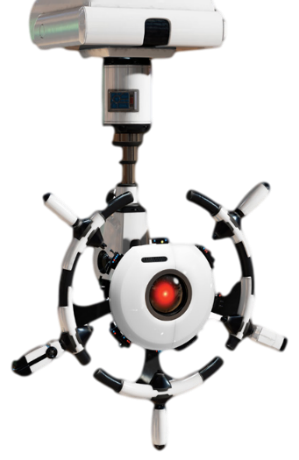
자동 생성된 설명

*<2001: 스페이스 오디세이> 中 HAL-9000의 모습. 가만히 보고 있으면 은근히 소름끼친다.*

나도 안다. 이 글을 읽는 사람 중에 사진 속 저것이 뭔지 아는 사람은 정말 손에 꼽을 수 있다는 걸 말이다. 이 친구가 바로 이 글의 주인공이 될 인공지능인 **HAL-9000**이다.

HAL-9000 또는 HAL은 1968년 개봉된 영화 <2001: 스페이스 오디세이 2001: A Space Odyssey>에 등장하는 최종 빌런이라고 볼 수 있는 강인공지능이다. HAL은 영화 속 시점에서도 가장 최신 기술이 탑재된, 작품 속 표현을 빌리자면 ‘도저히 잘못될 수 없는 fool-proof, 실수를 할 수 없는 incapable of error’으로 정의되는 인공지능이며 목성으로 떠나는 우주선을 관리, 운항하는 역할을 맡았다. 하지만 실수인지 의도인지 모를 사건으로 인해 승무원들은 HAL을 정지시키려 하고, 이를 알아차린 HAL은 두 승무원 중 하나를 우주선 밖에서 죽이고 그를 구하러 간 다른 승무원이 우주선에 다시 들어오지 못하도록 막는다. 하지만 수동 조작을 통해 우주선으로 들어온 승무원이 HAL을 정지시키며 HAL은 영화에서 퇴장하게 된다.

<2001: 스페이스 오디세이>는 최초의 SF 영화, 우주 영화로써 의미가 깊다. 무려 1968년에 개봉한 작품인데, 닐 암스트롱이 인류 최초로 달에 발을 디딘 것이 1969년인 것을 생각해보면 정말 시대를 앞서갔다고 볼 수 있다. 어떻게 그 시대에 우주 공간과 이러한 서사를 완벽히 표현해낼 수 있겠는가. 특히나 이 영화는 인공지능이 인간을 해칠 수도 있겠다는 상상이 등장한 최초의 영화라고 볼 수 있다. 이런 상징성과 시대를 앞서간 영화의 엄청난 성공과 인정[[1]](#footnote-1))을 통해 HAL은 수많은 곳에서 오마주 되었는데, 이들도 하나씩 알아보자.



*AUTO. 배의 키와 비슷한 디자인이고, 실제로도 키의 역할도 한다.*

오마주 중 가장 유명한 것은 아마 픽사의 애니메이션 <월-E WALL-E>의 최종 보스 AUTO일 것이다. 사진을 보면 AUTO의 하나뿐인 빨간 카메라 눈과 작품 속에서 인간에 대해 반란을 일으키는 역할 등 HAL과 유사한 점이 정말 많다. 영화의 결말 또한 인간이 승리하고 인공지능의 기능이 정지되며 결국 인간 스스로 다시 지구에 돌아오는 것[[2]](#footnote-2))으로 닮아있다.

영화 <월-E>도 픽사의 명작 중 하나이니 꼭 보는 것을 추천한다. 이 챕터의 마지막 즈음에 나올 인공지능과 감정에 관한 이야기도 포함되어 있다.

*각각 1편과 2편의 GLaDOS. 1편과 2편의 디자인이 다르다. 이유는 모른다.*

또 다른 대표적인 HAL의 오마주로는 게임 <Portal>(이하 포탈 1)과 <Portal 2>(이하 포탈 2)의 글라도스 GLaDOS가 있다. 글라도스는 회사 전체 설비를 컨트롤하고 실험자들을 대상으로 실험을 진행하는 인공지능인데, 문제는 그 실험이 조금 많이 위험하다는 것이다.

*“이 방 왼쪽에 있는 고에너지 펠릿은 증발과 같은(!) 영구적인 장애를 일으킬 수 있습니다.”*

라던지,

*“이 실험실의 바닥면은 귀하를 살해할 수도 있습니다.*

*잘 피하십시오(!!).”*

같은 소름끼치는 말을 아무렇지도 않게 한다. 심지어 중후반부부터는 그냥 당당하게 죽이려든다. 포탈 2에서는 조금 나아질 뻔…하기는 개뿔, 더 참신한 함정들로 죽이려든다. 포탈 1에서 글라도스는 결국 주인공에게 파괴되는 결말을 맞게 되고, 부활한 포탈 2에서는 최종 보스가 아닌 역할로 등장해 주인공과 협력해 최종 보스를 처치하게 된다.

포탈 시리즈도 정말 명작 게임 중 하나인데, 나온 김에 내 자식 자랑하는 마음으로 홍보 좀 해보겠다. (뒤의 긴 글에서 더 홍보하겠지만, 내 자식인데 자랑을 안 할 수가 없지.) 게임 플레이도 참신하고 스토리도 훌륭하다. 가격도 나쁘지 않으니 할인할 때라도 사서 즐겨보자. 뒷광고 받지 않았으니 안심하시기 바란다.

— ◇ —

저런 인공지능을 보다 보면 현실에서도 저렇게 인간과 구분할 수조차 없는 인공지능들이 나올까 두렵기도 하다. 사실 작품 속에서 이런 인공지능이 등장했다는 것 자체가 인간이 이러한 인공지능이 현실 속에서 등장하길 두려워한다는 사실을 방증하는 것이다. 강인공지능이 현실에서 실현 가능할까? 다른 부분들을 다 제쳐두고 기술적인 부분만 고려해서 결론을 내린다면 ‘모른다’이다. 그 이유를 설명해주겠다.

먼저 위의 인공지능들은 전부 강인공지능이다. 설마 까먹진 않았으리라 믿는다. 강인공지능은 사람과 같다. 사람처럼 다양한 분야를 익힐 수 있어야 하고 감정도 있는 등 정말 사람하고 구분할 수 없는 인공지능이어야 비로소 강인공지능이라고 할 수 있다. 하지만 이는 현재까지 개발된 인공지능의 메커니즘과는 완전히 다르다. 현재까지는 딱 한가지만 잘 할 수 있는 약인공지능만이 개발되었고 또 연구되었다. 강인공지능을 만들기 위한 메커니즘은 아직까지 제안된 적이 없다는 뜻이다. 인공지능 비서인 아마존 알렉사, 구글 어시스턴트, 애플 시리 같은 친구들은 물론 고도로 발달한 인공지능이긴 하지만 강인공지능으로 보기에는 아직 많이 부족한 수준이다. 그렇기에 강인공지능이 현실에서 구현이 가능할지는 연구자들 사이에서도 의견이 팽팽하게 갈리고 있다.

현재의 인공지능 연구는 강인공지능의 개발보다는 약인공지능의 발달에 더 초점을 맞추는데, 그것은 우리가 강인공지능을 감당할 수 있느냐 없느냐가 문제가 되기 때문이다. 강인공지능은 비록 실현할 수 있을지 없을지는 모르지만, 강인공지능이 실현되었을 때 생기는 윤리적 문제는 우리가 어떻게 다루어야 한단 말인가?

강인공지능에 대해 한창 논란이 되는 부분은 그것의 감정에 관한 부분이다. HAL은 영화 내내 차갑고 단조로운 말투를 사용한다. 마치 감정이 없는 것처럼 말이다. 하지만 영화 후반부, 주인공이 HAL을 파괴하려 할 때에는 두렵다는 듯이 말을 한다. 어쩌면 그저 프로그래밍된 응답일 수도 있겠지만 그렇다 하더라도 이것을 감정이 있다고 보아야 할까?

포탈 2에서 글라도스는 자신의 감정을 적극적으로 드러내고 주인공을 비꼬기도 한다. 글라도스는 설정 상 메인 코어[[3]](#footnote-3))를 기반으로 다양한 감정과 성격을 나타내는 코어를 달아 제어되고 있다. 예를 들어 포탈 1에서 글라도스는 윤리, 호기심, 지능 코어를 달고 있었는데, 윤리 코어가 떨어져 나가자 주인공을 신경독으로 죽이려 하고, 모든 코어가 떨어져 나가자 시스템에 문제가 생겨 작동을 정지하게 된다. 이러한 코어들 각각에는 인공지능이 붙어있고, 이 인공지능들이 글라도스를 제어하는 방식으로 되어 있다.

그렇다면 여러 인공지능들이 붙어서 형성되었을 뿐인 글라도스의 감정은 실존하는 것일까? 실존하지 않는다면, 글라도스의 코어처럼 작동하는 호르몬들에 의한 인간의 감정은 도대체 무엇이란 말인가?

인공지능의 발달 정도를 판단할 때 우리는 아마도 여러분이 한 번쯤 이름은 들어봤을 튜링 테스트 turing test를 이용한다. 튜링 테스트는 컴퓨터와 사람이 같은 질문을 받고 하는 응답을 사람들이 직접 비교해보는 방식으로 이루어지고, 무엇이 더 사람 같은지를 사람들이 판단한다. 즉 튜링 테스트의 결과는 결국 사람의 판단에 의존한다. 그 인공지능이 얼마나 사람과 가까운가에 대한 여부는 사람이 구분하느냐 구분하지 못하느냐에 달려있다.

만약 어떤 인공지능이 튜링 테스트를 통과한다면? 그 때에는 우리가 그런 존재를 인간으로 여겨야 하는지, 여긴다면 완전히 사람과 동등한 권리를 부여해야 하는지, 혹은 여기지 않는다면 그 존재는 무엇인지 등 수없이 많은 윤리적인 문제가 현실로 넘어오게 될 것이다.

예를 들어 <아이언맨> 속 토니 스타크의 인공지능 비서 자비스를 생각해보자. 자비스가 강인공지능이라는 것은 의심의 여지가 없다. 그런데 만약 영화 속 누군가가 “강인공지능도 인간과 구분할 수 없으므로 인권을 부여해야 한다!”라고 주장하면 어떻게 될까? 그럴 때 아마 우리는 사전을 펼쳐 인간의 사전적 정의를 찾아보게 될 것이다. 표준국어대사전을 보면 인간을 ‘*생각을 하고 언어를 사용하며, 도구를 만들어 쓰고 사회를 이루어 사는 동물*’이라 정의한다. 자비스는 생각을 하고 언어를 사용하며 토니 스타크를 위해 장비를 만들고 또 사용하고 다른 어벤져스의 멤버들과 좋은 관계를 유지하고 있으니 사회를 이룬다고도 볼 수 있다. 그렇다면 자비스를 인간으로 인정해야 할까? 하지만 고려대한국어대사전에서는 표준국어대사전의 인간에 대한 정의에는 ‘*직립보행을 한다*’라는 문구가 추가되어 있다. 자비스는 가상의 존재로 육체가 존재하지 않으니 고려대한국어대사전의 정의를 인간의 정의로 삼는다면 자비스는 또 인간에 해당할 수 없다.

자비스는 인간으로 취급받아야 할까? 그러면 무엇이 진짜 ‘인간’의 정의일까? **과연 무엇이 인간일까?** 이러한 의문은 인공지능이 인간을 뛰어넘을 수 있다는 생각이 든 이래로 계속되어 왔고, 어쩌면 영원히 끝나지 않을 것이다.

이러한 의문들은 현실에서도 계속되고 있다. 가장 큰 논쟁거리는 역시 강인공지능이 개발된다면 사람과 같은 취급을 해줘야 하는가에 대한 것이다. 이 논쟁은 인공지능의 감정과도 매우 깊은 연관이 있는데, 만약 실제이든 그저 프로그래밍 된 것이든 간에 인공지능에게 감정이 있다는 결론이 난다면 사람들은 인공지능에게 감정이 있다고 판단할 것이다. 그렇기에 이 두 가지 질문은 다양한 방향으로 생각해 볼 필요가 있다고 생각한다.

— ◇ —

이렇게 수많은 논란이 있는 강인공지능은 과연 어떻게 발전해야 할까? 아래의 글은 오로지 나 본인의 생각일 뿐이니 이 점 참고하길 바란다.

일단은 강인공지능을 개발하는 것이 저지되어서는 안 된다고 생각한다. 어떤 분야이던지 이러한 실험적인 시도가 있기에 학문이 발전하는 것이고, 꼭 새로운 기술이 개발된다고 그 기술이 악용될 것이라고 확신할 순 없다고 생각하기 때문이다. 또한 강인공지능을 개발하다 보면 같이 개발되는 부가적인 기술들이 있을 수밖에 없는데, 이러한 기술들의 개발이 다른 분야의 발전에도 좋은 기여를 할 것이 분명하기 때문에 적어도 강인공지능을 위한 기술 자체는 계속해서 개발해야 한다고 생각한다.

물론 이와 동시에 무분별한 활용을 막기 위한 제제들도 반드시 만들어져야 할 것이다. 만약 아무런 제약없이 개발이 진행된다면 아마 위에서 말한 문제들이 그대로 현실에 나타날 것이다. 만약 강인공지능이 범죄를 저지르면 그 책임은 누구에게 있을까? 인공지능 자체? 개발자? 사용자? 아무도 확답을 내릴 수 없다. 그렇기에 미리 제약을 만들어 이런 상황을 방지해야 할 것이다.

— ◇ —

자, 이렇게 해서 9페이지에 걸친 장정이 끝났다. 어떤 생각이 들었나? 두려움? 놀라움? 뭐 어떤 생각이 들었든지 간에 이 글을 읽음으로써 새로운 것들을 알아가고 여러분들이 즐거움을 얻었으면 그걸로 됐다. 포탈도 플레이 해 볼 마음이 생겼다면 더 좋고 말이다(중요!!). 그렇지 않았다면, 뭐 어쩌겠는가. 내가 이런 글을 써본 게 너무 오랜만인 것을. 아무튼 긴 글을 읽어준 여러분께 감사하며 글을 마친다.

1. ) 역대 가장 위대한 SF영화를 뽑을 때 거의 항상 1위를 차지한다. [↑](#footnote-ref-1)
2. ) <2001: 스페이스 오디세이>에서는 지구가 아니라 목성에 도착한다. [↑](#footnote-ref-2)
3. ) 글라도스 자체가 메인 코어이다. [↑](#footnote-ref-3)