인문논총 제78권 제4호 (2021.11.30), pp. 93~125
[DOI] https://doi.org/10.17326/jhsnu.78.3.202111.93

인공지능과 문예적 창의성

— 허구적 상상력을 중심으로*

윤 주 한**

[초 록]

본고는 '인공지능이 창의성을 갖출 수 있을까'라는 질문을 적절하게 다루는 하나의 방식을 제안하기 위하여 기획되었다. 이 질문에 답하기 위해 가장 먼저 해야 할 일은 창의성의 개념을 명료화 하는 것이다. 나는 창의성을 '새로운 산물을 산출하는 체계적인 인과적 기능을 수행하는 심적 능력'으로 정의하고, 우리가 심적 능력에 대한 기능주의를 고려해볼만한 철학적 상정으로 인정한다면, 이를 디딤돌 삼아 인공지능과 창의성에 대한 고찰을 시작해볼 수 있다고 주장한다.

다음으로, 예술 창작 인공지능의 사례로서 현재 가장 발전된 자연어 처리 인공지능들 중 하나인 GPT-3의 현주소를 다룬다. GPT-3는 방대한 양의 학습과 매개변수를 바탕으로 (부분적으로는) 인간의 것과 거의 구분하기 어려울 정도의 글을 생산해낸다. 그러나 GPT-3는 여전히 문예 '작품'을 생산해낼 수 있는 수준의 창의성에는 이르지 못했다.

* 이 논문은 서울대학교 인문학연구원이 지원한 집담회의 성과임.
** 대구대학교 자유전공학부 조교수

주제어: 창의성, 인공지능, 문예 창작, 허구적 상상력, GPT-3
Creativity, Artificial Intelligence, Creative Writing, Fictive Imagination, GPT-3

이러한 관찰을 토대로 본고는 인공지능이 문예 작품을 생산해낼 수 있는 수준의 창의성에 이르기 위해서는 어떤 능력이 필요한지를 검토한다. 나는 문예적 창의성을 갖추기 위해서는 '허구 세계'를 (재)구성하고 재현하는 능력, 즉 허구적 상상력이 전제되어야 함을 논증하고, '가능세계 상자' 모델을 통해 허구적 상상의 메커니즘을 밝힌다.

1. 서론

퀴렐 교수는 입을 벌리고 앉아 있는 다른 학생들을 향해 천천히 몸을 돌렸다.

"이 괴물이 무엇인지 알고 있나요?"

갑작스런 침묵 속에서 해리가 말했다. "선생님이죠?"

"그렇지 않아요,"라고 퀴렐 교수는 말했다. 그의 입술이 일그러졌다. "음모죠."

당황스러운 침묵이 흘렀다.[1]

소설 <해리 포터> 시리즈의 한 부분처럼 보이는 이 구절은 인공지능 프로그램인 GPT-3에 의해 작성된 것이다. 일본에서는 인공지능이 쓴 소설이 문학상의 예선을 통과했다.[2] 인공지능이 그린 초상화가 크리스티 경매에서 43만 2,500달러에 낙찰되는 일도 벌어진다.[3]

우리의 예술 관행에서, 인공지능 예술은 더 이상 먼 미래의 이야기

1) Gwern Branwen (2021), "GPT-3 Creative Fiction" <https://www.gwern.net/GPT-3#harry-potter-and-the-methods-of-rationality> [accessed 7 October 2021].

2) 윤희일(2016), 「인공지능이 쓴 소설, 문학상 1차 심사 통과」, 경향신문, 2016. 03.22. <https://www.khan.co.kr/world/japan/article/201603220915521>.

3) 곽노필(2018), 「인공지능 그림 첫 경매…5억원에 팔렸다」, 한겨레, 2018.10.26. <https://www.hani.co.kr/arti/science/technology/867532.html>.

가 아니다. 한마디로 우리는 이미 인공지능이 예술작품을 '창작'하는 시대에 살고 있다. 그러나 '인공지능이 예술작품을 창작하는 시대'라는 구절은 여러 철학적 문제들을 제기한다. 먼저 인공지능이 '예술작품'을 창작하는 것이 맞는가? 즉, 인공지능이 어떤 이미지나 문자열을 생산할 수는 있겠지만, 그것을 '예술'로 분류하는 것이 타당한가? 혹은 '인공지능'이 예술작품을 창작하는 것이 맞는가? 즉, 예술작품을 창작하는 주체는 인공지능이 아니라 인공지능을 만든 사람이 아닐까? 혹은 인공지능이 예술작품을 '창작'하는 것이 맞는가? 즉, 인공지능이 작품을 제작하는 행위를 할 수는 있겠지만, 그것이 인간 예술가들이 하는 것과 마찬가지의 '창작행위'가 맞는가?

이러한 문제들은 각각 별도의 논의들을 필요로 하지만, 이 논의들의 기저에 있는, 혹은 이 논의들 전반과 모두 연관되는 질문이 있다. 그것은 바로 '인공지능이 (인간 예술가와 마찬가지로) 창의적일 수 있는가'에 관한 질문이다. 창의적이라는 수사가 당연히 예술이나 예술가들에게 독점적인 것이 아니지만, 적어도 '예술가들은 일반적으로 창의적이다'라는 주장은 널리 받아들여져 왔다. 만일 우리가 예술을 창조하는 '사람'들을 창의적이라고 부른다면, 예술을 (혹은 예술처럼 보이는 무언가를) 창조하는 '기계'는 어떤가? 이들을 과연 창의적이라고 부를 수 있을까? 혹은, 이들을 왜 창의적이라고 불러서는 안 되는가?

본고의 목적은 '인공지능은 (예술적으로) 창의적일 수 있는가'라는 질문을 정당하게 다루는 하나의 방식을 제안하는 것이다. 철학의 영역에서 인공지능, 그리고 창의성이라는 두 주제 모두 상대적으로 최근에야 주목되기 시작했으며, 특히 창의성이라는 개념이 갖는 모호성으로 인해 인공지능이 창의적일 수 있는가라는 질문에 대해서는 면밀한 철학적 논의가 거의 이루어지지 못했다. 그러나 앞서 언급했듯이 우리가 인공지능 예술이 제기하는 여러 철학적 문제들을 다루고자 한다면, 이 질문을 거쳐 가지 않을 수 없다.

인공지능의 창의성의 문제는 과학자들과 철학자들의 긴밀한 협업이 요구되는 주제들 중 하나이다. 인공지능이 무언가를 창조해낼 수 있게 하는 것은 과학자들의 몫이지만, 그것이 창의적인지를 판단하고, 더 나아가 창의적 인공지능이 지향해야할 방향을 제시하는 것은 철학자의 일이다. 본고는 바로 그러한 철학적 작업을 하기 위해 기획되었다.

더 나아가 인공지능이 창의적일 수 있는지의 문제를 다루는 것은 반대로 창의적이라는 것이 무엇인지에 대한 통찰을 얻기 위한 하나의 실마리가 될 수 있다. 창의성과 같은 광범위하고 모호한 용어(blanket term)를 정의하기란 쉽지 않다. 우리는 마치 깜깜한 방안에서 코끼리를 더듬듯, 우리 손에 잡히는 어떤 것을 우선 관찰할 필요가 있을지도 모른다. 본고는 창의성이라는 거대한 개념 중 인공지능과 관련된 부분을 찾아 더듬으려는 시도이다. 그 결과는 창의성이라는 코끼리의 전체 모습을 추정할 수 있는 힌트가 될 수 있다.

본격적인 논의에 들어가기 전에 이 연구의 범위에 대해 명시할 필요가 있다. 먼저, 이 연구는 '예술적' 창의성에 대해 다룬다. 창의성은 우리의 여러 관행에서 사용되는 포괄적 용어이다. 우리는 (혹은, 인공지능은) 창의적인 예술가일 수도 있고, 창의적인 바둑 기사일 수도 있고, 창의적인 경영자일수도 있다. 그러나 그 모든 관행에서 창의적이라는 속성이 부여되는 대상이 가지는 공통점들을 찾는 것은 간단한 일이 아니다. 나는 예술에서의 창의성에 집중함으로써, 바둑에서 창의적인 것과 예술에서 창의적인 것이 어떻게 같고 어떻게 다른지를 설명하는데 너무 많은 노력을 기울이지 않으려고 한다.

다음으로, 본고는 '문예적' 인공지능에 초점을 맞춘다. 그 이유는 크게 두 가지이다. 첫째, 나는 인공지능 예술 일반에 있어서의 창의성에 대해 이야기하기 전에, 서로 다른 예술 장르에서 활동하는 인공지능 예술가들의 창의성에 대해 별도로 살펴볼 필요가 있다고 본다. 즉, 인공지능과 창의성에 대한 논의는 상향식(bottom-up)으로 이루어지는 것

이 가장 적절하며, 나는 그 기저를 이루는 한 부분을 다루고자 한다. 둘째, 지금까지 인공지능 예술에 대한 연구는 주로 시각예술을 중심으로 이루어져왔다. 그러나 시각예술의 구성요소와 문학예술의 구성요소는 서로 다르기에, 시각 예술을 창조하는 인공지능의 창의성에 대한 분석이 문학예술에까지 일반화되지 않을 수 있다. 더 나아가, 인공지능 문학예술은 인공지능 시각 예술 이상으로 흥미로운 사례와 질문들을 제시한다.

이러한 점들을 염두에 두고 본격적인 논의를 시작해보자.

2. 창의성이라는 (모호한) 개념

인공지능이 창의적인지, 혹은 창의적일 수 있는지에 대한 질문에 답하기 위해 가장 먼저 해야 할 일은 그 질문 안에 포함되어 있는 개념들을 명확히 하는 것이다. 우리는 인공지능이 무엇인지에 대해서는 비교적 분명하게 알고 있다. 마빈 민스키(Marvin Minsky)에 따르면, 인공지능이란 "인간이 수행한다면 지능을 필요로 할 법한 일들을 수행하는 기계"이다.[4)5)] 그렇다면 창의성은 어떤가? 앞서 언급했듯이, 창의

4) 민스키의 원래의 정의는 학문(science)으로서의 인공지능에 대한 정의였다. 즉, 학문으로서의 인공지능이란 "인간이 수행한다면 지능을 필요로 할 법한 일들을 수행하는 기계를 만드는 학문"이다. Marvin Minsky (1968), "Preface", in *Semantic Information Processing* (ed. by Marvin Minsky), MA: MIT Press, p. v.

5) 반면 인공지능의 정의가 여전히 불명확하며 이에 대한 깊이있는 논의가 요청된다는 주장 역시 제기된 바 있다. Pei Wang (2019), "On Defining Artificial Intelligence", *Journal of Artificial General Intelligence*, 10.2, pp. 1-37. 인공지능의 엄밀한 정의에 대한 추가적인 심도 깊은 논의에 대해서는 다음을 참고하라. Dagmar Monett and others (2020), "Special Issue "On Defining Artificial Intelligence" —Commentaries and Author's Response", *Journal of Artificial General Intelligence*, 11.2, pp. 1-100.

성은 여러 관행에서 사용되는 포괄적 용어이다. 적어도 철학에서 창의성에 대하여 합의된 정의는 아직까지 없다고 봐도 무방할 것이다.[6] 더 나아가 마거릿 보든(Margaret Boden)은 창의성에 대한 본질주의적 접근이 오히려 우리를 오도할 수 있다고 주장한다.[7] 다만 그는 자율성, 지향성, 의식, 가치의 담지, 감정 등 우리가 창의적이라고 부르는 대상들이 경향적으로 소유하고 있는 속성들을 귀납적으로 찾아볼 수 있다고 부연한다.[8]

나는 창의성에 대한 본질주의적 정의의 가능성을 부정하지는 않지만, 이것이 우리가 당장 성취할 수 있는 목표가 아니라는 점을 인정한다. 또한 창의성 일반의 본성에 대해 최종적이고 확정적인 답변을 제시하는 것은 본고의 범위를 넘어선다. 따라서 내가 취할 전략은 우리가 인공지능의 창의성의 문제에 답하기 위해 꼭 필요한 정도로 창의성이라는 개념의 윤곽을 드러내는 것이다.

이를 위해 우리가 먼저 고려해야 하는 것은 '창의성'이라는 속성이 행위자와 산물, 혹은 과정 중 어느 쪽에 부여되어야 하는지의 문제이다. 우리는 어떤 사람이 창의적이라고 흔히 말하는 한편, 그가 생산한 산물을 두고 창의적이라고 말하기도 하고, 어떤 사람이 산물을 생산하는 과정을 창의적이라고 말하기도 한다.[9] 이 세 입장은 서로 배타적이지 않다. 즉, 창의성을 산물에 귀속되는 속성으로서 다루면서도 창의성이 행위자에게 귀속될 때를 고려할 수도 있고, 창의적 과정에 대해 논하면서도 그 산물의 창의성에 대해 논할 수도 있다. 혹은, 창의성을

6) Margaret A. Boden (2014), "Creativity and Artificial Intelligence: A Contradiction in Terms?", in *The Philosophy of Creativity*, New York: Oxford University Press, p. 226.

7) Margaret A. Boden (2014), p. 226, 233.

8) Margaret A. Boden (2014), pp. 233-236.

9) Dustin Stokes (2016), "Imagination and Creativity", in *The Routledge Handbook of Philosophy of Imagination*, Routledge, pp. 267-281 (p. 247).

이 세 가지에 동시에 적용되는 것으로 다루는 것도 가능하다. 다만 창의성과 관련하여 행위자, 산물, 과정 중 어느 쪽에 주목하는지에 따라 논의의 방향이 조금씩 달라질 수 있다. 우리가 여기에서 주목하고 있는 문제는 '인공지능이' 창의적일 수 있는지이므로, 본고에서는 행위자에게, 더 정확히는 행위자의 심적 능력에 부여되는 속성으로서의 창의성에 주목할 것이다.

그러나 우리 논의의 맥락에서 이러한 규정은 곧 중요한 도전에 부딪히게 된다. 창의성이 심적 능력이라면, 과연 인공지능이 창의성을 가질 수 있는가? 만일 창의성이 심적 능력이고, 심적 능력을 인간만이 가지는 특성으로 정의한다면, 인공지능의 창의성에 대한 논의 자체가 무의미하게 될 것이다.

보든은 창의성과 관련하여 중요하게 다루어지는 요소로서[10] 자율성, 지향성, 의식, 가치의 담지, 감정 등을 검토한 후 인공지능의 창의성에 대한 논의를 다음과 같이 결론 짓는다.

> 컴퓨터가 "정말로" 창의적일 수 있는지에 대한 질문은 현재 답변될 수 없다. 왜냐하면 그것은 [자율성, 지향성, 의식 등과 같은] 몇 가지 매우 논쟁적인 철학적 질문을 포함하기 때문이다. 만약 우리가 자율성에 관한 논증을 진지하게 받아들인다면, 컴퓨터가 일반적으로 추정되는 것보다 훨씬 더 독립적일지라도, 우리는 "인공지능-창의성"이 모순어법이라는 것에 동의할 수 있다. 그러나 우리가 지향성이나 의식에 호소한다면, 그 질문은 열려 있어야 한다.[11]

---

10) 앞서 언급했듯이, 보든은 창의성에 대한 본질주의적 접근을 거부하므로, 이 요소들이 창의성에 필수적(necessary)이라거나, 본질적(essential)이라거나, 창의성의 구성 요소(constitutive feature)라는 등의 표현을 사용하지 않는다. 그는 단지 우리가 창의적이라고 부르는 대상들에서 이러한 요소들이 경향적으로 발견된다고 주장한다.

11) Margaret A. Boden (2014), p. 242.

자율성에 관한 논증이란, 인공지능은 인간에 의해 '프로그램(programmed)' 되므로 자율적이지 못하고, 따라서 창의적일 수 없다는 논증이다. 그러나 보든 역시 적절하게 지적하듯이, 동시대의 딥러닝 기술은 컴퓨터가 프로그램된 것의 범위를 벗어난 어떤 결과물을 제시할 수 있도록 만들었다.[12] 혹자는 지금의 인공지능 프로그램 역시 인공지능이 일련의 과업을 수행할 때 반드시 준수해야 할 일종의 규칙이나 제약을 여전히 부여한다고 반박할 수 있다. 그러나 그렇다고 하더라도 이것이 '인공지능의 창의성'이라는 용어를 모순어법으로 만드는 것은 아니다. 왜냐하면 마치 시조가 엄격한 정형화된 규칙 내에서 창작되듯이, 혹은 연극이 시간과 공간의 제약 등과 같은 매체적 제약을 가지듯이, 인간 역시 규칙과 제약 내에서 창작활동을 한다고 볼 수도 있기 때문이다. 물론 우리는 그러한 규칙과 제약을 깬 예술가들을 상대적으로 '더 창의적'이라고 말하지만, 그렇다고 해서 다른 예술가들이 규칙과 제약 내에서 창작한다는 이유만으로 그들이 전혀 창의적이지 않다고 말하지는 않는다.

그렇다면 지향성이나 의식과 같은 심적 상태의 문제는 어떨까? 즉, 인공지능이 지향성이나 의식과 같은 심적 상태를 가질 수 있을까? 보든이 제안하듯이[13], 현대의 심리철학과 심리학, 인지과학 등에서 널리 받아들여지고 있는 기능주의(functionalism)는 심적 상태나 심적 능력이 인간만의 전유물이 아니라는 점을 시사한다.

기능주의에 따르면 한 심적 상태란 특정한 체계적인 인과적 기능(systemic causal function)을 수행하는 상태이다.[14][15] 가령 고통(pain)이

12) Margaret A. Boden (2014), pp. 229-232.

13) 보든이 정확하게 여기에서 제시하는 것은 계산주의 이론(computational theory)이다. 그러나 계산주의는 기능주의의 일종이며, 지향성이나 의식과 같은 심적 상태가 인간만의 전유물이 아니라는 주장은 기능주의 일반에서 발견되는 주장이기에, 여기에서는 기능주의로 치환하여 사용할 것이다.

14) 기능주의에 대한 일반적 설명에 관해서는 다음을 참고하라. Janet Levin (2018),

란 신체를 구성하는 한 조직에 대한 손상을 입력값으로 받아 주체가 그 손상의 원인이 되는 것을 피하게끔 하거나, 그 고통을 어떤 식으로든 표명하게 하는 등의 체계적인 인과적 기능을 수행하는 심적 상태라고 할 수 있을 것이다. 이때 중요한 것은 기능주의는 그러한 심적 상태의 기반이 되는 내적 구조가 무엇인지에 대해서는 문제 삼지 않는다는 것이다. 어떤 존재가 고통을 느낀다고 할 때, 기능주의에게 있어서 중요한 것은 그 고통이 수행해야하는 바로 그 체계적인 인과적 기능을 수행하고 있는지의 여부이지, 그 고통이 인간의 두뇌와 신경망에 의해서 발생하는지, 외계인이 가진 미지의 신경체계에 의해 발생하는지, 아니면 실리콘 칩으로 구성되어 있는 컴퓨터 CPU와 마이크로소프트 윈도우 운영체제에 의해 발생하는지는 중요하지 않다(기능주의의 일종인 계산주의(computationalism)와 연결주의(connectionism)가 인공지능 연구의 기반이 되는 것은 바로 이 때문이다).

나는 기능주의가 경쟁 이론들에 비해 우리의 심적 상태나 심적 능력의 본성을 가장 잘 설명할 수 있는 이론이라고 주장하고자 하는 것이 아니다. 심적 상태나 심적 능력의 본성이 궁극적으로 무엇인가를 논의하는 것은 본고의 범위를 벗어나는 일이다. 다만, 우리에게는 인공지능이 창의적일 수 있는지에 대한 물음을 유효한 것으로 다루기 위한 하나의 철학적 상정이 필요하고, 기능주의가 바로 그러한 상정이 될 수 있다는 것이다. 적어도 우리가 기능주의를 고려해볼만한 철학적 상정으로 인정한다면, 우리는 그런 상정을 디딤돌 삼아 인공지능과 창의성에 대한 고찰을 시작해볼 수 있다.

더 나아가 기능주의가 제안하는 것처럼 우리가 창의성을 체계적인

"Functionalism", *The Stanford Encyclopedia of Philosophy,* <https://plato.stanford.edu/archives/fall2018/entries/functionalism/>.

15) 기능주의를 설명할 때 일반적으로 '체계적인'이라는 수식어가 포함되지는 않지만, 본고의 이후 논의를 통해 그 필요성이 드러날 것이다.

인과적 기능으로 보는 것이 가지는 또 다른 유용성도 있다. 즉, '체계적인 인과적 (기능)'이라는 단서 조항은 거트(Berys Gaut)가 걱정하는 것과 같이 우연히 창의적으로 '보이는' 산물을 만들어낸 사람에게 (혹은 인공지능에게) 창의적이라는 속성을 부여하는 상황을 배제할 수 있다.[16] 가령 납치를 당해 구속복으로 묶여있는 사람이 흰 벽이 있는 방에 갇혀서 몸부림 치다가 우연히 옆에 있는 페인트통을 넘어뜨렸고, 그 페인트를 온 몸에 뒤집어 썼다고 해보자. 그는 구속복을 벗기 위해 몸부림치면서 수십 번 벽에 몸을 부딪쳤다. 그런데 우연하게도 그가 벽에 몸을 부딪치면서 만들어낸 페인트자국이 무척이나 아름다운 추상회화 같은 모양을 만들어냈다. 추측컨대 우리는 그것만으로 이 사람이 창의성을 지녔다고 말하지 않을 것이다. 대신에, 우리는 그가 이 산물을 '어떻게' 만들어 냈는지를 물을 것이다.[17] '체계적', '인과적'이라는 단서 조항은 바로 그러한 '어떻게'를 설명해준다. 어떤 기능이 체계적으로 인과적이기 위해서는 어떤 산물 x가 바로 그 기능에 의해 산출된 것이며, 모든 조건이 동일할 때 그 기능은 x와 동일한, 혹은 유사한 수준의 산물을 생산해 낼 수 있을 것이라고 합리적으로 기대될 수 있어야만 한다. 우리는 예의 납치된 사람을 동일한 방에 동일한 방식으로 가두어 놓았을 때 그의 몸부림이 또 다시 아름다운 추상회화같은 것을 만들어 낼 것이라고 합리적으로 기대할 수 없다.

다음으로, 우리는 창의성이 '어떤' 체계적인 인과적 기능을 수행하는 심적 능력인지를 물어야 한다. 일반적으로 창의성은 새로운(novel) 물질적, 정신적 산물을 산출하는 능력으로 간주된다. 다만 이때의 새

16) Berys Gaut (2009), "Creativity And Skill", in *The Idea of Creativity*, Brill, pp. 83-103.

17) 한편 거트는 바로 이 때문에 창의성을 정의할 때 그 산물 뿐만 아니라 '행위자(agent)'를 고려해야 한다고 주장한다. 이와 관련된 국내의 논의는 다음을 참고하라. 임수영(2020), 「거트의 창의성 이론에 대한 비판적 이해—창의성(Creativity)과 기술(Skill)의 관계를 중심으로」, 『미학』, 86.3, pp. 131-168.

로움의 정체를 무엇으로 볼 것인지에 대해서는 다양한 논의가 존재한다. 우선 '누구에게' 새로운 것인가? 보든은 창의성을 새로움의 정도에 따라 심리적 창의성(P-creativity)과 역사적 창의성(H-creativity)으로 구분한다.[18] 어떤 산물이나 아이디어가 창조자 자신에게만 새로워도 심리적 창의성을 부여받을 수 있다면, 역사적 창의성의 경우 창조자 자신뿐만 아니라 인간의 역사에서 새로울 것을 요구한다. 다만 이해완이 타당하게 지적하듯이, 새로움이 역사적으로 최초인지, 아니면 개인에게 최초인지는 창의성을 정의하는데 있어서 근본적으로 중요한 문제가 아닐 수 있다.[19] 베이즈 정리라는 것이 있는지도 몰랐던 열 살 아이가 바로 지금 이 정리를 스스로 발견했다고 해보자. 이 발견은 우리에게 이미 알려져 있는 것이기 때문에 학술적 가치를 거의 가지지 않는다고 말할 수는 있겠지만, 우리에게 이 정리가 알려져 있다는 사실만으로 이 아이의 창의성을 베이즈의 창의성보다 낮게 평가해야만 하는 것은 아닐 수 있다.

다음으로, 새로움이 긍정적인 가치를 가지는 것이어야 하는지 아닌지의 문제 역시 논쟁적이다. 가령 독재국가의 어떤 교도관이 지금까지 그 누구도 생각하지 못했던, 더 효과적으로 더 큰 고통을 줄 수 있는 새로운 고문 방법을 고안했다고 해보자. 우리는 이 사람을 창의적이라고 해야 할까, 아닐까? 창의성이 가치중립적 개념이라고 주장하는 사람들은 이 사람을 여전히 창의적이라고 여길 수 있겠지만, 창의성이 긍정적 가치를 가지는 것이어야만 한다고 주장하는 사람들은 이 사람이 창의적이지 않다고 말해야할 것이다.

이러한 논쟁점들에도 불구하고 여기에서 중요한 것은 창의성이 적

18) Margaret A. Boden (2004), *The Creative Mind: Myths and Mechanisms*, Routledge, p. 43.

19) 이해완(2021),「결과에서 품성으로: 창의성의 가치에 대한 개념적 분석」,『제147회 목요 콜로키움 발표문』, 서울대학교 예술문화연구소.

어도 ‘어떤 새로운 것’을 창조해내는 능력이라는 점에 대해서는 거의 합의가 이루어진 것으로 보인다는 점이다. 새로움이 역사적으로 새로워야 하는지 아닌지, 혹은 새로움이 긍정적인 가치를 가져야만 하는지 아닌지 역시 중요한 문제이지만, 적어도 여기에서 우리가 초점을 맞추어야 할 주제는 아니다. 일단 창의성이 ‘새로운 것’을 산출하는 심적 능력이라는 개념을 얻어낸 것에 만족하자.

요컨대 우리는 창의성을 새로운 산물을 산출하는 체계적인 인과적 기능을 수행하는 행위자의 심적 능력으로 본다는 최소한의 합의에 도달했다.[20] 이것을 창의성의 핵심 정의(Core Definition, CD)로 부르자. 이 정의는 여전히 창의성의 개념을 어느 정도 모호한 채로 남겨두지만, 우리가 인공지능의 창의성의 문제를 논할 때 최소한으로 필요한 자리채우기(placeholder)의 역할은 충분히 할 수 있을 것으로 보인다. 그렇다면 이제 우리에게 남은 문제는 인공지능이 그러한 체계적인 인과적 기능을 수행할 심적 능력을 가질 수 있는지이다.

3. 인공지능 예술의 현주소: GPT-3를 중심으로

서론에서 언급했듯이, 이미 많은 인공지능 ‘예술가’들이 작품을 생산하고 있다. 인공지능 예술의 가장 눈부신 발전은 시각 예술의 영역에서 이루어지고 있다. 우리는 인공지능이 만들어낸 이미지들을 보면서 놀라고, 경탄하며, 심지어 미적 경험을 하기도 한다. 그러나 앞서의 납치된 사람과 흰 벽에 묻은 페인트의 사고실험에서 볼 수 있듯이, 시

20) 한 가지 염두에 두어야 할 점은 이 능력을 예술가나 천재 발명가와 같은 일부의 사람들만이 가진 것으로 볼 필요는 없다는 것이다. 우리가 모두 근력을 가지고 있지만 사람에 따라 그 정도가 다르듯이, 우리 모두가 어느 정도 창의성이라는 능력을 가지고 있지만 사람에 따라 그 탁월성의 정도가 다르다고 보는 편이 더 적절할 것이다[이해완 역시 유사한 주장을 펴고 있다. 이해완 (2021).].

각적 이미지가 우연적인 조합에 의해서 새롭게 보이는 어떤 것이 될 수 있는 경우는 비교적 쉽게 생각해볼 수 있는 반면, 상당한 길이의 문자열이 우연하게도 말이 되면서(making sense) 새롭게 보이는 어떤 것이 되기는 더 어렵다. 따라서 창의성의 문제에 관해서라면 시각 예술의 사례보다 오히려 문학예술의 사례에서 더 흥미로운 통찰을 얻을 수 있을지 모른다.

그렇다면 인공지능은 문학 예술을 창작할 수 있을까? 인공지능으로 문학작품을 저술하고자 했던 최초의 시도는 1984년에 발표된 <그 경관의 턱수염은 반쯤 만들어졌다>(The Policeman's Beard Is Half Constructed)로 거슬러 올라간다.[21] 이 산문, 혹은 시집은 'Racter'라는 이름의 인공지능에 의해 만들어졌다고 알려졌는데, 사실상 임의적으로 산출된 영어 문장을 조합하는 수준에 그쳤다.[22] 이후 멕시카(MEXICA) (1999), 브루투스(BRUTUS) (2000) 등과 같은 이야기 생성 프로그램들이 개발되었지만, 그다지 주목할만한 성과는 없었다는 것이 중론이다. 그러나 심층학습(deep learning) 기술이 인공지능에 본격적으로 적용되면서 인공지능 기술은 '코페르니쿠스적 전환'을 이루었다. 바둑을 두는 인공지능 '알파고(AlphaGo)'가 심층학습을 기반으로 하는 대표적 인공지능 중 하나이다.

심층학습을 적용하면서 자연어 처리(Natural Language Processing) 인공지능의 수준 역시 빠르게 높아졌다. 특히 GPT-3는 현재(2021년 10월) 시점에서 가장 높은 수준의 언어적 능력을 갖추었다고 여겨지는 자연어 처리 인공지능들 중 하나이다.[23] 2020년 6월에 발표된

21) Chamberlain and Racter (1984), "The Policeman's Beard Is Half Constructed", NY: Warner Books.

22) Leah Henrickson (2021), "Constructing the Other Half of The Policeman's Beard", *The Electronic Book Review*.

23) 현재 Open AI의 GPT-3 외에도 구글의 '람다(LaMDA)', 화웨이의 '판구 알파(PanGu Alpha)', 네이버의 '하이퍼클로바' 등의 자연어 처리 인공지능들이 개발

GPT-3는 Open AI에서 개발한 자연어 처리 특화 인공지능 GPT (Generative Pretrained Transformer)의 3세대 버전이다. GPT-3는 그 전 세대인 GPT-2와 동일한 모델을 사용하지만, 훨씬 방대한 양의 텍스트를 학습하며 GPT-2의 100배 이상인 1,750억 개의 매개변수(parameter)를 가진다.[24]

GPT-3의 연구자들과 사용자들은 GPT-3가 기존의 자연어 처리 인공지능들에 비해 괄목할만한 발전을 이루었다고 보고한다. GPT-3를 검토한 플로리디와 치리아티(Floridi and Chiriatti)는 GPT-3로 생산한 글이 적어도 부분적으로는 인간의 글과 거의 분간하기 어려울 정도의 수준에 이르렀다고 주장한다.[25] GPT-3로 글을 생산하는 실험을 수행한 미국의 작가이자 연구자 그웬 브란웬(Gwern Branwen)은 GPT-3가 단지 인간의 수준에 근접했을 뿐만 아니라 GPT-3의 산물은 "창의적이며, 위트 있고, 깊이 있으며, 메타적이고 때로는 아름답다"고 주장한다.[26][27]

GPT-3가 우리에게 주는 놀라움에도 불구하고 GPT-3의 능력에 대한 여러 의심들이 제기될 수 있다. 먼저, GPT-3는 새로운 작품을 창조한다기 보다는 기존의 작품을 응용하여 이어 쓰는 수준에 그치고 있다.

되어 공개된 바 있다. 심지어 람다, 판구 알파, 하이퍼클로바 모두 GPT-3보다 더 많은 매개변수를 가진다. 그럼에도 불구하고 이들은 현재로서는 GPT-3에 비해 연구와 검증이 충분히 이루어지지 못한 상황이다.

24) Tom B. Brown and others (2020), "Language Models Are Few-Shot Learners", *ArXiv Preprint ArXiv:2005.14165*, p. 8.

25) 하지만 그들의 최종적 결론은 GPT-3가 수학적, 의미론적, 윤리적 측면 모두에서 결함을 가진다는 것이다. Luciano Floridi and Massimo Chiriatti (2020), "GPT-3: Its Nature, Scope, Limits, and Consequences", *Minds and Machines*, 30.4, pp. 681-694.

26) Gwern Branwen (2021).

27) 분량의 문제로 GPT-3가 생산한 산출물들에 대한 자세한 소개는 생략한다. 산출물의 샘플들은 Gwern Branwen (2021)에서 확인할 수 있다.

더 구체적으로 말해서, GPT-3는 전체로서의 작품을 만든다기 보다는 작품처럼 보이는 '조각'을 만드는 수준이다. 다음의 글은 본고의 서두에서 인용한 GPT-3의 글에서 이어지는 부분이다. GPT-3는 레스 롱(Less Wrong)이 저술한 <해리 포터> 시리즈의 팬픽션(fanfiction) 작품인 <Harry Potter and the Methods of Rationality>의 일부 구절을 입력값으로 받고, 그 뒤를 이어 이 글을 썼다.

> 퀴렐 교수는 입을 벌리고 앉아 있는 다른 학생들을 향해 천천히 몸을 돌렸다.
> "이 괴물이 무엇인지 알고 있나요?"
> 갑작스런 침묵 속에서 해리가 말했다. "선생님이죠?"
> "그렇지 않아요,"라고 퀴렐 교수는 말했다. 그의 입술이 일그러졌다. "음모죠."
> 당황스러운 침묵이 흘렀다.[28]

> 그러자, 후플푸프 소녀는 "그 이야기는 일리가 있고, 저는 엄청나게 불길한 예감을 느껴요."라고 소리쳤다.
> 그녀의 옆에 앉아 손을 들고 있는 소녀는 헤르미온느 그레인저였다. "음, 그것은 이 장(chapter)에 이야기가 있기 때문이에요."
> 그리고 그 이후 강의는 완전히 중단되었다.
> "그러므로 만약 여러분이 디멘터 한 마리보다 산악괴물을 더 강하게 만드는 것이 무엇인지 궁금하다면, 그것은 한쪽이 눈에 보이고 다른 쪽은 보이지 않아서가 아니라, 산악괴물이 더 강력하고 더 지능적이며 더 공격적이기 때문입니다. 하지만 디멘터의 일은 단순히 가장 큰 영혼을 가진 사람을 찾아내는 것이기 때문에, 디멘터가 물리칠 수 있는 강력하고 지적인 생명체도 있습니다."[29]

28) Gwern Branwen (2021).
29) Gwern Branwen (2021).

각각의 문장은 말이 되고, 그럴듯하며, 심지어 소설 <해리 포터> 시리즈의 한 구절들처럼 보이기도 하지만, 그 전체로서는 일관된 서사를 이루지 못한다. GPT-3가 영국 <가디언>지에 '기고'했다는 칼럼 역시 하나의 단일한 글이라기 보다는 여러 조각들을 편집한 것에 지나지 않는다. (다만, <가디언>의 편집자는 GPT-3의 글을 편집하는 것이 인간 기고자의 글을 편집하는 것과 별반 다를 것이 없었다고 주장한다.[30]) 다음으로, GPT-3는 그 개발자들이 'prompt'라고 명명한, 인간 사용자의 지시, 혹은 선입력에 의해서만 작동한다. 즉, 인간 예술가가 창작의 시작과 끝을 자율적으로 결정하는 것과는 대조적으로, GPT-3는 인간 예술가가 방아쇠를 당겨주지 않으면 아무것도 생산하지 않는다. 마지막으로, GPT-3는 그 원리상 단지 '확률 게임'을 하고 있을 뿐이다. 즉, GPT-3는 단지 어떤 단어나, 문장, 혹은 문구 다음에 왔을 때 통계적으로 가장 적합할 확률이 높은 단어들을 순차적으로 선택할 뿐이다. 물론 GPT-3는 방대한 학습량과 1750억 개에 달하는 세밀한 매개변수(parameter)를 통해 우리에게 무척 성공적으로 보이는 게임을 수행하지만 말이다.

그러나 이러한 의심들 중 어느 것도 GPT-3와 창의성 사이의 관계를 결정적으로 단절시킬 수 있을 것으로는 보이지 않는다. 앞서의 의심들에 대해 각각 살펴보자. 먼저 GPT-3가 단지 기존의 것을 응용할 뿐이라고 하더라도 그것이 GPT-3를 창의적이라고 부르지 못할 타당한 이유가 되지는 않는다. 톰 스토파드의 희곡 <로젠크란츠와 길덴스턴은 죽었다>는 <햄릿>의 응용이지만 그 이유 때문에 스토파드가 창의성이지 않다고 말하지는 않는다. 다음으로, (인간 사용자의) 선입력이 필요하다는 것 역시 GPT-3가 창의적이지 않다는 주장을 타당하게 뒷받침해주는 이유가 될 수 없다. 정몽주의 <단심가>는 이방원의 <하여가>

30) GPT-3 (2020), "A Robot Wrote This Entire Article. Are You Scared yet, Human?", *The Guardian*, 2020 <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>.

라는 '선입력'에 대한 답가였다. 만일 누군가 제인 오스틴의 미완성 소설 <샌디턴(Sanditon)>을 이어서 그 소설을 훌륭하게 완성한다면, 그를 창의적이라고 부르지 못할 이유는 없어 보인다. GPT-3가 단지 통계적으로 단어들을 고를 뿐이라는 의심은 앞서의 두 가지에 비해 더 근본적인 문제 제기가 될 수 있다. 그러나 실상 우리는 인간이 어떠한 원리로 말하거나 쓸 단어들을 선택하는지에 대해 명확히 알지 못한다. 우리가 어떤 단어 다음에 올 가장 적절한 단어를 찾는 방식이 (GPT-3와 마찬가지로) 학습의 결과로 통계적으로 최적의 단어를 찾는 것일 수도 있다. 더욱이 우리가 기능주의를 채택한다면, GPT-3가 어떤 메커니즘에 의해 새로운 산물을 만들어 내는지 별로 중요하지 않을 수 있다. 그 메커니즘이 특정한 인과적 기능을 수행할 수 있다면 말이다.

우리는 현재 인공지능이 할 수 있는 것과 할 수 없는 일을 냉정하게, 그리고 정확하게 파악할 필요가 있다. 즉, GPT-3를 통해 자연어 처리 인공지능이 곧 인간의 수준에 오를 것이라는 장미빛 전망에 빠지는 것이 적절하지 않은 것과 마찬가지로, GPT-3의 창조물에 대해 기대할 것이 전혀 없다는 강경한 회의주의 역시 적절하지 않다. 분명 GPT-3는 하나의 완결된 작품을 창작하지 못한다는 점에서 가장 초보적인 소설가의 수준에도 도달하지 못했다. 그러나 GPT-3가 만들어낸 글의 '일부'는 새로울 뿐만 아니라 명망있는 소설가의 글만큼이나 유려하고 눈길을 끈다.

그렇다면 GPT-3는 과연 창의적이라고 말할 수 있는가? 이 질문을 적절하게 다루기 위해서는 먼저 우리가 창의성을 정확히 어떤 의미로 '사용할지'를 물어야만 한다. CD로 돌아가 보자. CD에 따르면, 창의성이란 새로운 산물을 산출하는 체계적인 인과적 기능을 수행하는 심적 능력이다. 앞서 언급했듯이 CD는 여전히 어느 정도의 모호함을 가지고 있지만, 우선 우리의 목표가 창의성에 대한 최종적이고 확정적인 정의를 갖는 것이 아니라는 점을 다시금 상기하자. 또한 우리는 범용

인공지능(Artificial General Intelligence)에 대해 논하고 있는 것도 아니다. 범용 인공지능의 개발은 인공지능 연구의 최종적인 목표일 수 있지만, 근미래에 성취될 수 있는 목표는 아니다. 우선은 바둑두기, 이미지 만들기, 이미지에 캡션 달기, 글쓰기 등 특정한 과업을 제대로 수행할 수 있는 인공지능을 개발하는 것이 현재 우리가 서 있는 단계이다. 그렇다면 새로움이나 창의성의 개념 역시 모든 인간사에 적용되는 것일 필요가 없을 것이다. 지금 우리는 문예창작 능력을 갖춘 인공지능에 대해 논하고 있으므로, 여기에서는 문예적 영역에서 새로운 산물을 산출하는 체계적인 인과적 기능에 대해 논하는 것으로 충분하다.

그렇다면 문예적 창조, 특히 서사적인 허구적 문예작품의 창조에서 새로움이란 무엇인가? 서사적 허구 문예작품을 '이야기하기', 즉 '스토리텔링(storytelling)'으로 보았을 때, 우리는 새로움을 크게 두 가지 종류로 나눌 수 있다. 즉, '스토리'에 있어서의 새로움과 '텔링'에 있어서의 새로움이다. 스토리란 이야기의 내용으로서, 등장인물들과 사건들을 포함한 '허구적 세계(fictional world)'를 전제한다. 즉, 하나의 완결된 스토리를 창작한다는 것은 한 허구 세계를 창조한다는 것이다. 다음으로 텔링이란 이야기의 기법, 혹은 양식(스타일)이다. 우리는 스타인벡의 무미건조하지만 통렬한 문체에 대해, 이청준의 자기 침잠적인 문체에 대해 경탄한다. 이러한 분석을 바탕으로 서사적 문예작품의 창의성을 크게 세 단계로 구분해보자.

1) **A-창의성 단계**: 학습된 내용을 바탕으로 주어진 허구 세계 내에서 응용된 작품을 만드는 단계
2) **W-창의성 단계**: 학습된 내용을 바탕으로 새로운 허구 세계를 창조하는 단계.
3) **S-창의성 단계**: 학습된 내용을 바탕으로 새로운 문예 양식(style)을 발명하는 단계.

GPT-3의 성과로 미루어볼 때, GPT-3는 A-창의성 단계에 근접해가고 있는 것으로 보인다. 그러나 GPT-3가 아직 A-창의성 단계에 도달했다고는 볼 수는 없다. GPT-3는 응용은 할 수 있으나 '작품'을 만들지는 못하기 때문이다. 하나의 완결된 서사 작품을 만들기 위해서는 글의 각 요소들이 유기적으로 연결되어야 하고 정합적인 이야기 구조를 가져야 한다. 물론 포스트모더니즘 이후로 완결된 서사라는 개념이 많은 비판을 받은 것은 사실이지만, 그것은 반대로 문학예술이 완결성을 갖는 서사구조를 가져야 한다는 생각이 문학에 대한 전통적인 관념임을 시사한다. 인공지능 문학이 모두 포스트모더니즘적 해체 서사라고 주장하지 않는 한, 이 관념은 무시될 수 없을 것이다.

GPT-3가 A-창의성 단계에 도달하지 못했다는 것은 경험적으로 그것이 W-창의성과 S-창의성에도 도달하지 못했다는 것을 의미한다. W-창의성과 S-창의성은 A-창의성보다 더 어려운 과제이기 때문이다. 그러나 인공지능이 각 단계에 도달할 수 있을지를 검토해보는 것은 여전히 의미있다. 우선 S-창의성은 우리가 근미래에 도달하기에 어려운 과제인 것처럼 보인다. 시각예술에서의 양식이 비교적 분명하고 가시적인데 반해 문학예술에서의 양식은 보다 미묘하고(subtle) 불분명하다. (인공지능 시각예술의 양식에 대한 연구는 진행되고 있지만 인공지능 문학예술에서의 양식에 대한 연구는 거의 없는 것 역시 유사한 이유 때문으로 보인다.[31]) 더구나 양식은 지속되는 정체성을 가진 일종의 인격을 전제한 개념이다.[32] 인공지능이 어떤 양식을 흉내낼 수는 있겠으나, 인공지능이나 그 산물이 어떤 지속적이며 고유한 형식

31) 인공지능 시각예술에서의 양식에 관한 논의는 다음을 참고하라. Leon A. Gatys Alexander S. Ecker and Matthias Bethge (2015), "A Neural Algorithm of Artistic Style", ArXiv Preprint ArXiv:1508.06576.

32) 이에 대해서는 리처드 볼하임(Richard Wollheim)의 '개인적 스타일(individual style)'에 관한 논의를 참고하라. Richard Wollheim (1987), "Pictorial Style: Two Views", in *The Concept of Style* (ed. by Berel Lang), Cornell University Press.

적 특징을 가질 수 있는지, 그리고 그 지속적이고 고유한 형식적 특징을 단순한 반복이 아닌 양식이라고 볼 수 있는지에 대해 더 많은 논의가 필요하다.

그렇다면 W-창의성은 어떨까? 사실 A-창의성과 W-창의성은 함께 다루어져야만 하는데, 왜냐하면 응용된 것이든 새로운 것이든 하나의 완결된 작품을 만든다는 것은 하나의 작품 세계(work world)를 구성한다는 것을 뜻하기 때문이다. 작품의 완결성, 혹은 정합성은 단순히 개별 단어와 단어, 개별 문장과 문장 사이의 연결이 자연스러운지에 달려 있는 것이 아니라, 그것이 하나의 정합적이고 완결적인 허구 세계를 바탕으로 하고 있는지에 달려있다. 모든 정합적인 허구 서사작품은 정합적인 허구 작품 세계(fictional work world)를 갖는다. 허구 작품 세계란 그 작품에 대한 허구적 참들로 구성된 세계이다.[33] 월튼에 따르면, 허구작품이라는 소도구(prop)는 발생의 원칙(principle of generation)에 의거해 허구적 참들을 발생시킨다.[34] 그리고 이러한 허구적 참들을 도출해낼 수 있는 능력이 바로 허구 세계를 (재)구성하고 재현하는 능력이다. 그 세계가 하나의 완결된 허구작품에 의해 주어진 것이든, 아니면 하나의 주어진 '어떨까(what-if) 문장'을 가지고 구성한 것이든, 아니면 무(無)로부터 시작해 완전히 새롭게 만든 것이든, 어떤 허구 세계를 (재)구성하고 재현할 수 있는 능력이 있어야만 A-창의성 단계에도 완전하게 도달할 수 있다. 즉, A-창의성과 W-창의성은 함께 추구되어야 하는 목표이다.

다시 말해, 허구 세계를 구성하고 재현할 수 있는 능력은 적어도 서

33) 작품 세계의 개념은 월튼에게서 빌려온 것이다. 작품 세계에 대한 월튼의 논의는 다음을 참고하라. Kendall L. Walton (1990), *Mimesis as Make-Belive*, Harvard University Press.

34) 발생의 원칙이란 어떤 믿는체 하기 게임(make-believe game)에서 한 소도구가 허구적 참을 발생시키도록 하는 원리이다. 발생의 원칙은 이 게임에 참여하는 참여자들 간의 상호 이해에 의해 만들어진다. Kendall L. Walton (1990), p. 38.

사적 문예작품을 창작하는 창의성의 구성적 요소라고 볼 수 있다. 따라서 우리의 당면 과제는 인공지능이 허구 세계를 구성하고 재현할 수 있는 능력을 갖출 수 있을지를 검토하는 것이다.

요컨대, '인공지능은 창의적일 수 있는가'라는 질문은 그 자체로는 적절하게 다루어질 수 없다. 우리가 범용 인공지능을 다루는 것이 아니며, 창의성의 개념 역시 광범위하고 모호하기 때문이다. 이 질문을 정당하게, 그리고 유의미하게 다루기 위해서는 창의적 범용 인공지능으로 향해가는 길목에서 우리가 밟아갈 수 있는 중간 단계들을 찾아내야 한다. 지금까지의 분석을 통해 우리가 찾아낸 하나의 단계는 허구 세계를 구성하고 재현하는 능력을 갖추는 것이다.

그렇다면 우리가 이제 물어야 할 질문은 다음과 같이 서술될 수 있다.

'어떤 인공지능 x는 허구 세계를 구성하고 재현하기 위한 체계적인 인과적 기능을 수행하는 능력을 갖출 수 있는가?'

다음 장에서는 허구 세계를 구성하고 재현하기 위한 체계적인 인과적 기능을 수행하는 능력이 무엇인지, 그리고 인공지능이 그러한 능력을 갖출 수 있을지를 검토할 것이다.

4. 상상하는 인공지능

<해리 포터> 시리즈의 소설만을 읽은 사람 A와, 같은 시리즈의 영화만을 본 B가 서로 이야기를 하고 있다. A가 다음과 같이 말한다. "우리 지도교수가 덤블도어 같은 사람이면 좋겠어." 이에 B는 다음과 같이 답한다. "그는 덤블도어라기보다는 스네이프에 가깝지." 두 사람은 서로 다른 작품을 감상했음에도 불구하고 덤블도어와 스네이프가

같은 세계에 속하는 허구적 캐릭터라는 사실을 알고 있다. 즉, 두 사람은 허구적 작품 세계에 대한 관념을 가지고 있다.

허구 작품을 창조하는 것뿐만 아니라, 허구 작품을 감상하는 것에도 허구 세계에 대한 관념, 혹은 허구 세계를 (재)구성하고 재현하는 심적 행위, 혹은 능력이 필요하다. 그리고 우리는 이러한 행위를 상상하기, 이러한 능력을 '상상력'이라고 불러왔다.[35]

상상은 창조성만큼이나 포괄적인 개념이다. 상상이라는 개념이 적용되는 경우는 허구 작품의 세계를 구성하거나 재구성할 때뿐만 아니라, 특정한 음악의 선율을 머릿속에 떠올릴 때, 누군가에게 발생한 실제 이야기를 들으며 그 장면을 생생하게 떠올릴 때 등 다양하다. 이 중 허구 세계를 (재)구성하거나 재현할 때 발생하는 상상하기를 '허구적 상상(fictive imagining)'이라고 부른다.[36]

이제 우리는 앞선 장의 말미에서 도출된 질문의 답 중 일부를 얻었다. 허구 세계를 구성하고 재현하기 위한 체계적인 인과적 기능을 수행하는 능력은 바로 허구적 상상력이다. 그렇다면 허구적 상상은 정확히 어떻게 이루어지는가? 니콜스와 스티치(Shaun Nichols and Stephen Stich)는 '가능세계 상자(Possible World Box)' 모델을 통해서 우리의 인지 구조(cognitive structure) 내에서 허구적 상상이 일어나는 메커니즘을 설명하고자 한다.[37]

가능세계 상자 모델을 설명하기 위해서는 먼저 심리학과 최근의 심리철학에서 주로 받아들여지는, 우리가 믿음이나 욕구와 같은 명제태

35) 상상 능력이 창의성 일반과 밀접한 관계를 맺고 있다는 주장에 대해서는 다음을 참고하라. Dustin Stokes (2016).

36) 나는 다른 지면에서 허구적 상상의 본성에 대한 보다 자세한 논의를 다루었다. Juhan Yoon (2020), "A Theory of Fictional Art: Issues on Nature, Value, and Media", Doctoral thesis, Seoul National University.

37) Shaun Nichols and Stephen Stich (2000), "A Cognitive Theory of Pretense", *Cognition*, 74.2, pp. 115-147.

도를 갖는 상태를 설명하기 위한 표준적인 개념적 틀에 대한 설명이 필요하다. 이 개념적 틀에 따르면 우리가 믿음이나 욕구를 갖는다는 것을 이해하는 최선의 방법은 우리의 인지 체계 내에 구별되는 기능적 역할을 하는 정신적 워크스페이스(workspace), 혹은 비유적으로 말해서 '상자(boxes)'가 작동하고 있는 것으로 이해하는 것이다. 이에 따르면, '김철수는 서울에 산다'는 믿음을 가지는 것은 그 믿음의 내용에 상응하는 개별 표상(representation token)이 우리의 '믿음 상자(Belief Box)'에 저장되는 것으로 가장 잘 이해될 수 있다. 마찬가지로 '저 케이크가 먹고 싶다'는 욕구를 가지는 것은 그 욕구의 내용에 상응하는 표상이 우리의 '욕구 상자(Desire Box)'에 저장되는 것으로 가장 잘 이해될 수 있다.

믿음 상자에 들어가는 대부분의 표상 토큰은 지각 프로세스(perceptual process)에 의해 생성된다. 즉, 우리가 외부 세계로부터 얻어낸 정보들은 표상 토큰으로서 우리의 믿음 상자 내에 저장되고, 이 표상 토큰들이 모여 실제 세계에 대한 우리의 표상을 이룬다. 한편 욕구 상자 속의 표상 토큰은 주로 우리의 신체 모니터링 시스템(body monitoring system)에 의해 생성된다. 이 표상 토큰들은 우리가 '바라는' 실제 세계의 모습에 대한 표상을 이룬다.

믿음 상자와 욕구 상자는 단지 표상들을 구획화하여 가두어 놓는 데서 그 역할이 끝나지 않는다. 믿음 상자와 욕구 상자 속의 표상들은 의사 결정 시스템(decision-making system)과 행동 조절 시스템(action control system)을 거쳐 궁극적으로 우리의 행동을 유발한다. (그림 1.)

믿음 상자와 욕구 상자는 신경학적으로 우리 뇌에 존재하는 영역은 아니지만, 우리가 어떤 명제 태도를 갖는 상태를 명료하게 이해할 수 있도록 돕는 개념 틀(framework)이다. 그러나 이 개념 틀에는 결정적으로 결여되어 있는 것이 있다. 즉, 믿음 및 욕구 상자를 상정하는 것만으로는 우리가 어떠한 '가장(pretense)'이나 '반사실적 상상(counterfactual

imagining)’에 참여하는 것에 대한 적절한 이해를 제시할 수 없다는 것이다. 예컨대, 내가 빈 컵을 가지고 커피를 마시는 척한다고 해보자. 이 가장을 성공적으로 수행하기 위해서는 ‘이 컵이 커피로 차 있다’에 해당하는 표상을 가지는 것이 필요하다. 그러나 나의 믿음 상자나 욕구 상자에 그에 해당하는 표상이 있을 리 없다. 만일 그런 표상이 나의 믿음 상자에 있었다면, 나는 커피를 마시는 척하는 것이 아니라 커피를 실제로 마시려고 시도하고 또 실패해야 할 것이다.

니콜스와 스티치는 가장하기를 수행하고 이해하는 우리의 능력을 가장 잘 설명하기 위해서는 우리가 믿음 상자나 욕구 상자와 같은 기능을 수행하는 별도의 워크스페이스, 즉 ‘가능 세계 상자(Possible Worlds Box)’를 가진 것으로 간주하는 것이 가장 타당하다고 제안한다.[38] 가능 세계 상자와 믿음/욕구 상자의 차이는 그 기능에 있다. 믿음 상자가 “세계가 어떠한지”에 대한 표상을 담고 욕구 상자가 “세계에 대해 원하는 바”에 대한 표상을 담는다면, 가능 세계 상자는 “우리가 참으로 믿지도 않고 참이 되기를 원하지도 않는 일단의 상정(assumption)들이 주어졌을 때 세계는 어떠할지”에 대한 표상을 담는 것으로 이해된다.[39] 예를 들어 ‘나는 슈퍼맨이다’와 같은, 가장을 위한 최초의 전제(initial premise)가 주어졌을 때, 비어있던 가능 세계 상자는 우리의 추론 메커니즘에 의해 그 상정된 세계에 대한 보다 상세한 기술들로 채워진다. 이에 따라 가능 세계 상자에서 표상된 세계는 더 구체적이 되고 우리는 진행되고 있는 가장하기에 적합한 행동들을 효과적으로 선택할 수 있게 된다.

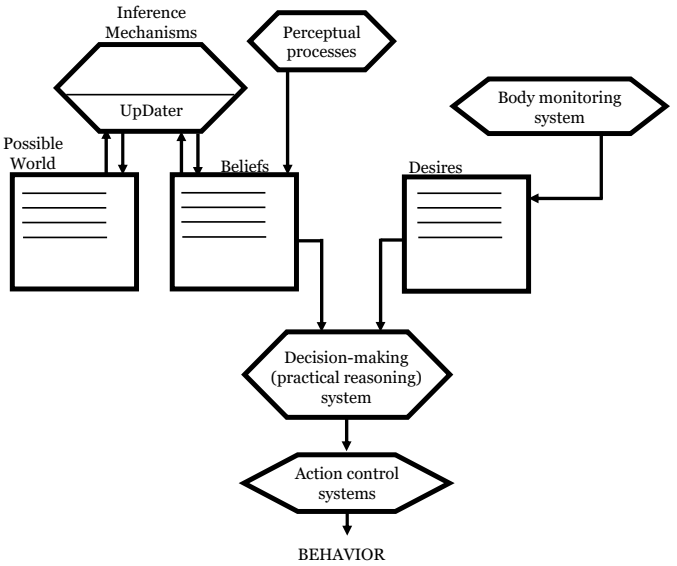
가능 세계 상자 내의 표상들은 (상자에 의해) 믿음 상자나 욕구 상자의 표상들과 구획화(compartmentalized)됨으로써 허구 세계에 대한 우리의 관념이 실제 세계에 대한 믿음이나 욕구와 엉망으로 뒤섞이는

38) Shaun Nichols and Stephen Stich (2000).

39) Shaun Nichols and Stephen Stich (2000), p. 122.

것을 막을 수 있다. 그러나 더욱 중요한 것은 가능세계 상자 내의 표상들이 우리의 인지 체계 내에서 고립(isolated)되지는 않는다는 것이다. 만일 가능 세계 상자가 믿음/욕구 상자와 고립되었다면, 우리가 최초의 전제만을 가지고 지속적으로 가장하기를 수행할 수 있다는 점을 설명할 수 없을 것이다. 가령 '나는 아이언맨이다'라는 최초 전제를 부여받은 어린아이는 이 가장하기에 적합한 행동들을 지속적으로 지시받지 않더라도 대개 스스로 적절한 행동들을 선택하고 수행해 나간다. 또한 우리는 어떤 허구 작품을 감상하면서 그 세계의 모든 부분에 대한 정보가 주어지지 않더라도 실제 세계에 대한 우리의 기존의 믿음들을 통해 그 부분들을 채워나간다.

이러한 고립을 막는 메커니즘이 바로 '업데이터(UpDater)'이다. 니콜스와 스티치는 '업데이터'라는 추론 메커니즘이 우리의 믿음 상자와 가능 세계 상자 속의 표상들을 적절하게 조정하는 기능을 수행한다고 설명한다. 가령 '지금 비가 온다'라는 표상이 우리의 믿음 상자에 들어왔을 때 업데이터는 '오늘 우산을 가져갈 필요가 없다'라든지 '오늘은 세차하기 좋은 날이다'와 같이 새로운 믿음과 배치되는 믿음들을 제거하거나 변경한다. 한편 '나는 슈퍼맨이다'와 같은, 가장하기를 위한 최초 전제가 나의 가능 세계 상자에 들어온다면, 업데이터는 나의 믿음 상자에 있는 표상들 중에 슈퍼맨인 체 하는 가장하기에 적합하지 않은 표상들을 제외한 나머지 표상들을 모조리 가능 세계 상자로 집어넣는다. 이러한 방식으로 가능 세계 상자 속에 표상된 세계는 주어진 가정들을 제외하고 우리가 가진 실제 세계에 대한 표상과 동일하게 된다.



[그림 1] '가능세계 상자 모델'의 알고리즘[40]

가능세계 상자 모델은 가장하기와 관련된 우리의 일반적 심적 상태나 능력을 설명하는 모델이지만 허구적 상상하기 역시 잘 설명한다. 가능세계 상자는 반사실적 상정(counterfactual assumption)이 주어졌을 때 작동한다. 우리가 허구작품을 감상할 때, 우리는 일련의 반사실적 상정들을 읽고 이해하며, 가능세계 상자는 이 상정들이 실제 세계에 대한 우리의 믿음들과 뒤섞이지 않도록 구획화하는 기능을 한다. 우리가 허구작품을 창작할 때, 가능세계 상자는 반사실적 상정들을 차례로 수립하면서 한 허구 작품 세계를 구성하기 위한 토대를 세우고, 업데이터는 믿음 상자 속의 표상 토큰들이 이 상정들과 정합적으로 어울리도록 조정하면서 이 작품 세계를 정합적으로 채워나간다. 즉, 가능

40) Shaun Nichols and Stephen Stich (2000), p. 123.

세계 상자 모델은 허구작품을 창작하고 감상할 때 우리 마음 속에서 일어나는 일들을 적절히 설명해줄 수 있는 이론적 모델이다.

요컨대 문예적 창의성을 갖추기 위해서는 허구 세계를 구성하고 재현하기 위한 체계적인 인과적 기능을 수행하는 능력, 즉 허구적 상상력을 갖추어야 하며, 허구적 상상하기는 가능세계 상자 모델을 통해서 적절하게 설명될 수 있다.

정리하자면, 우리는 인공지능이 창의적일 수 있는지에 대한 물음을 적절하게 다루기 위한 한 방식으로서 그 물음을 인공지능이 문예적 창의성을 가질 수 있는지에 대한 물음으로 좁혔다. 또한 나는 문예적 창의성의 구성적 요소로 허구 세계를 (재)구성하고 재현하는 능력을 제시했고, 이 능력을 갖추는 것이 인공지능이 문예적 창의성을 갖추기 위한 선결 조건이라고 주장했다. 앞선 논의를 통해 이 능력이 곧 허구적 상상력이라는 점이 도출되었고, 우리는 허구적 상상을 설명하기 위한 이론적 모델을 살펴보았다.

그렇다면 우리에게 남은 과제는 과연 인공지능이 가능세계 상자 이론이 제안한 것과 같은 허구적 상상 메커니즘을 실제로 구현할 수 있을지를 규명하는 것이다. 아쉽게도 이 질문에 대한 결정적인 답변은 후속 연구로 미루어야 할 것 같다. 이 질문에 대한 결정적인 답변을 제시하기 위해서는 경험과학과의 보다 적극적인 협업이 필요하기 때문이다. 다만 가능세계 상자 모델이 향후 문예적 창의성을 갖춘 자연어 처리 인공지능의 개발 방향을 설정하는데 도움이 되는 몇 가지 실마리를 던져 줄 수 있다고 생각한다.

첫째, 인공지능이 허구 세계를 정합적으로 구성하기 위해서는 실제 세계를 정합적으로 이해(구성)하는 믿음 체계를 갖추어야 한다. 허구 세계를 허구적으로 만드는 것은 한정된 수의 가정적 상정인 반면, 그 사이를 채우는 것은 실제 세계에 대한 무수히 많은 우리의 믿음이기 때문이다. 둘째, 인공지능이 실제 세계를 정합적으로 구성할 수 있도

록 하기 위해서, 실제 세계를 정합적으로 구성하기 위한 핵심(core) 믿음들이 어떤 것인지, 그리고 이 믿음들을 인공지능이 가진다는 것은 어떤 것인지에 대한 연구가 필요하다. 현재의 심층학습 기술의 발전 방향으로 미루어볼 때 우리가 세계를 관찰하면서 세계에 대한 믿음을 축적하고 또 조정해나가는 것처럼, 인공지능 역시 그렇게 세계에 대한 일군의 믿음을 갖추는 것이 가능해 보인다. [문제는 인간의 미세 조정(fine tuning) 없이 얼마나 제대로 된 믿음을 축적할 수 있는지가 되겠지만 말이다.] 그렇게 하기 위해서는 단지 인터넷 상의 정보를 긁어모으는 것보다는 더 양질의 데이터셋(dataset)이 필요할 것이다. 셋째, 인공지능이 실제 세계를 정합적으로 이해하는 믿음 체계를 갖추었다면, 그 믿음 체계와 가능세계 상자 사이의 소통을 매개하는 업데이터의 구체적인 알고리즘이 필요하다. 앞서 이야기했듯이, 실제 세계에 대한 믿음을 적절하게 이용하지 않고서는 허구 세계를 (재)구성할 수 없기 때문이다.

이러한 점들이 해결되고 실현된다면, 우리는 허구적 상상력을 갖춘 인공지능에 한 발 더 가까워질 것이다.

5. 결론

인공지능은 빠른 속도로 발전하고 있으며 인공지능 예술은 그 여러 발전 방향 중 하나이다. 그러나 인공지능 예술은 인공지능 개발에서 결코 지엽적인 문제가 아니다. 예술은 인간의 고유한 능력들이 조화를 이루어 만들어내는 인간적 성취의 정점으로 여겨지기 때문이다. 만일 인공지능의 목표가 인간과 최대한 가까워지는 것이라면, 예술, 혹은 더 나아가 탁월한 예술을 창작하는 인공지능은 그야말로 인공지능 개발의 최종장이라고도 볼 수 있다.

나는 지금까지 '인공지능이 창의적일 수 있는가'라는 질문을 적절하게 다룰 수 있는 한 가지 방법에 대하여 논했다. 나의 주장은 인공지능이 창의성 일반을 갖출 수 있는가라는 질문은 너무 추상적이고 광범위하므로 이 질문을 특정 인공지능이 특정 분야의 창의성을 갖출 수 있는지에 대한 질문으로 구체화해야 한다는 것이다. 이렇게 했을 때 우리는 가령 문예적 인공지능이 문예적 창의성을 갖추기 위해서는 어떻게 해야 하는지를 물을 수 있다. 나는 문예적 창의성의 구성적 요소로서 허구 세계를 구성하고 재현하는 능력, 즉 허구적 상상력을 제안했다. 또한 우리가 허구적 상상에 참여하는 과정에 대한 하나의 설득력 있는 이론적 모델을 설명하고, 이것이 인공지능 연구에 던져주는 함의에 대해서 논했다. 이러한 논의들을 바탕으로 인공지능 기술이 더욱 발전한다면 우리는 인공지능이 인간과 얼마나 가까워졌는지를 파악하기 위해 앨런 튜링의 테스트, 즉 이미테이션 게임(imitation game)[41] 대신 기계가 허구적 상상을 할 수 있는지를 테스트하는 '이매지네이션(상상) 게임(imagination game)'을 필요로 하게 될 지 모른다.

그럼에도 불구하고 우리에게는 여전히 많은 과제들이 남아있다. 문예적 창의성과 인공지능의 관계에 대한 나의 논의가 다른 예술 장르나 예술 일반, 혹은 창의성 일반에 어떻게 적용될 수 있는지는 흥미로운 주제이다. 나는 여기에서 문예적 창의성과 인공지능의 관계로 논의를 제한했지만, 이 논의가 궁극적으로는 예술적 창의성 일반, 더 나아가 창의성 일반으로 확장될 가능성은 열려있다. 또한 앞선 장의 말미에서 언급했듯이, 허구적 상상을 수행하는 인공지능을 실제로 어떻게 구현할 것인지, 다시 말해, 가능세계 상자 모델을 튜링 기계나 인공 신경망의 개발에 어떻게 적용할 것인지 역시 경험과학과의 보다 적극적인 협업을 통해 탐구되어야 할 문제이다. 더 나아가, 만일 허구적 상상

41) Alan Turing (1950), "Computing Machinery and Intelligence", *Mind*, LIX.236, pp. 433-460.

력을 갖춘 인공지능이 등장한다면, 이 인공지능은 단지 완결된 허구 작품을 만드는 것뿐만 아니라 거짓말을 하거나 배우처럼 연기를 하는 것 역시 가능하게 될 것이다. 이것이 일으키는 인식론적, 윤리학적 문제 역시 철학적 고찰을 필요로 한다.

참고문헌

【자 료】

곽노필(2018),「인공지능 그림 첫 경매…5억원에 팔렸다」, 한겨레, 2018.10.26. <https://www.hani.co.kr/arti/science/technology/867532.html>

윤희일(2016),「인공지능이 쓴 소설, 문학상 1차 심사 통과」, 경향신문, 2016.03.22. <https://www.khan.co.kr/world/japan/article/201603220915521>

【논 저】

이해완(2021),「결과에서 품성으로: 창의성의 가치에 대한 개념적 분석」,『제147회 목요 콜로키움 발표문』, 서울대학교 예술문화연구소.

임수영(2020),「거트의 창의성 이론에 대한 비판적 이해—창의성(Creativity)과 기술(Skill)의 관계를 중심으로」,『미학』, 86.3.

Margaret A. Boden (2014), “Creativity and Artificial Intelligence: A Contradiction in Terms?”, in *The Philosophy of Creativity*, New York: Oxford University Press.

_____ (2004), *The Creative Mind: Myths and Mechanisms*, Routledge.

Gwern Branwen (2021), “GPT-3 Creative Fiction” <https://www.gwern.net/GPT-3#harry-potter-and-the-methods-of-rationality> [accessed 7 October 2021].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and others (2020), “Language Models Are Few-Shot Learners”, *ArXiv Preprint ArXiv:2005.14165.*

Chamberlain and Racter (1984), “The Policeman’s Beard Is Half Constructed”, NY: Warner Books.

Luciano Floridi and Massimo Chiriatti (2020), “GPT-3: Its Nature, Scope, Limits, and Consequences”, *Minds and Machines*, 30.4.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge (2015), “A Neural Algorithm of Artistic Style”, *ArXiv Preprint ArXiv:1508.06576.*

Berys Gaut (2009), “Creativity And Skill”, in *The Idea of Creativity*, Brill.

GPT-3 (2020), “A Robot Wrote This Entire Article. Are You Scared yet, Human?”, *The Guardian*, 2020 <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>.

Leah Henrickson (2021), “Constructing the Other Half of The Policeman’s Beard”, *The Electronic Book Review.*

Janet Levin (2018), “Functionalism”, *The Stanford Encyclopedia of Philosophy* <https://plato.stanford.edu/archives/fall2018/entries/functionalism>.

Marvin Minsky (1968), “Preface”, in *Semantic Information Processing* (ed. by Marvin Minsky), MA: MIT Press.

Dagmar Monett, Colin W. P. Lewis, Kristinn R. Thórisson, Joscha Bach, Gianluca Baldassarre, Giovanni Granato, and others (2020), “Special Issue “On Defining Artificial Intelligence”—Commentaries and Author’s Response”, *Journal of Artificial General Intelligence*, 11.2.

Shaun Nichols and Stephen Stich (2000), “A Cognitive Theory of Pretense”, *Cognition*, 74.2.

Dustin Stokes (2016), “Imagination and Creativity”, in *The Routledge Handbook of Philosophy of Imagination* (Routledge).

Alan Turing (1950), “Computing Machinery and Intelligence”, *Mind*, LIX.236.

Kendall L. Walton (1990), *Mimesis as Make-Belive* (Harvard University Press)

Pei Wang (2019), “On Defining Artificial Intelligence”, *Journal of Artificial General Intelligence*, 10.2.

Richard Wollheim (1987), “Pictorial Style: Two Views”, in *The Concept of Style* (ed. by Berel Lang), Cornell University Press.

Juhan Yoon (2020), “A Theory of Fictional Art: Issues on Nature, Value, and Media”, Doctoral thesis, Seoul National University.

원고 접수일: 2021년 10월 12일
심사 완료일: 2021년 11월 1일
게재 확정일: 2021년 11월 3일

ABSTRACT

Artificial Intelligence and Literary Creativity

Yoon, Juhan*

This paper aims to propose a way to properly deal with the question of 'Can artificial intelligence have creativity?' Firstly, to answer this question, the concept of creativity is clarified. I define creativity as a mental ability to perform a systematic causal function to produce novel products or ideas, and argue that if we recognize functionalism as a philosophical assumption worth considering, we can start discussing artificial intelligence and creativity making this assumption a steppingstone.

Next, as an example of art-creating artificial intelligence, I scrutinize where exactly GPT-3, one of the most advanced natural language processing artificial intelligence programs, stands. GPT-3 produces writings that are (partly) almost indistinguishable from humans' due to vast amounts of learning and parameters. However, GPT-3 still has not reached the level of creativity to create literary 'works'.

Then, I examine what ability artificial intelligence needs in order to reach the level of creativity to create literary works. I argue that in order to have literary creativity, the ability to (re)construct and represent fictional work worlds, i.e., fictive imagining, is required, and elucidate the mechanism of fictive imagining through the 'Possible World Box' model.

* Assistant Professor, Department of Liberal Arts, Daegu University