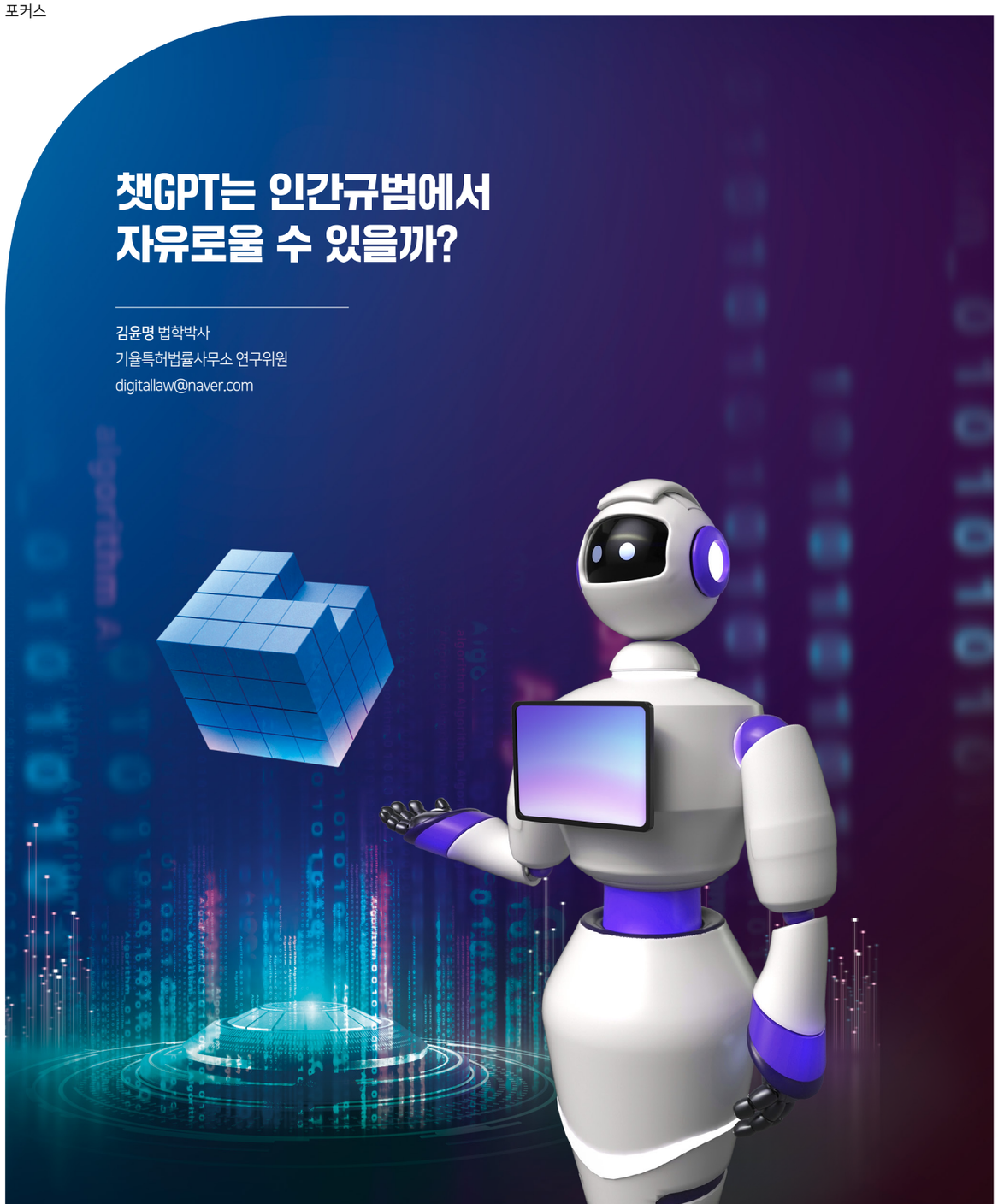
챗GPT는 인간규범에서 자유로울 수 있을까?

김윤명 법학박사
기울특허법률사무소 연구위원
digitallaw@naver.com



자연어를 학습한 생성형 언어모델이 등장하다

인공지능의 대중화를 가져온 알파고(AlphaGo) 이후, AI 분야의 혁신을 이끄는 챗GPT는 대중을 관중에서 선수로 끌어들이고 있다. 인공지능은 큰 그림을 심어줬지만, 실상 일상에서 가장 즐겨하는 수단으로 활용됨으로써 보다 현실적인 것으로 구현되고 있는 것이다. 챗GPT는 OpenAI에서 개발한 생성형(Generative) AI 모델로 인터넷에 공개된 대규모의 데이터를 바탕으로 기계학습해 만들어진 Generative Pre-trained transformer(GPT) model을 말한다. 인간의 언어로 필요한 내용을 입력하면 텍스트 유형에 상관없이 생성한다. 챗GPT에 프롬프트(Prompt) 형태의 명령어를 문장 또는 단어 등으로 입력하면 GPT가 명령어의 맥락을 분석하여 그에 맞는 결과를 생성한다. 그렇기 때문에 생성형 AI 모델을 말하는 GPT 이름이 붙었다. 그것도 채팅처럼 대화형으로 이뤄지므로 챗GPT라고 한다. 대화하듯 필요한 내용을 입력하고 검색 결과보다 쉽고 용이하게 완성된 결과물을 만들어주므로 선풍적인 인기를 끌고 있다.

챗GPT 현황은 놀랍다

챗GPT가 2022년 12월 출시된 이후, 5일 만에 100만 명 이상이 가입함으로써 다른 빅테크 서비스와 비교할 수 없을 정도로 확장세를 이어가고 있다. 챗GPT는 음성, 코드(에러 수정), 이미지, 텍스트, 경영 시뮬레이션 및 비디오 등 다양한 콘텐츠를 만들어내고 있다. 프롬프트(Prompt)에 원하는 '생각하는 내용'을 명령어처럼 입력하면 다양한 유형의 콘텐츠를 생성한다. 코딩 실력은 구글의 수억대 개발자의 능력을 넘어섰다고 한다. 컴퓨터공학에서 챗GPT 성능을 보고서 코딩교육 시스템을 바꿔야 한다는 얘기가 나온다. 기업에서는 한참 노코드(No code), 로우코드(Low code) 흐름이 이어졌는데, 이제는 챗GPT가 코딩해주기 때문에 코딩 인력이 필요 없는 상황에 직면하기도 한다. 따라서 인력양성도 바뀌어야 한다. 이제는 학교 교육에서 코딩 실습의 형태가 바뀌어야 할지도 모른다.

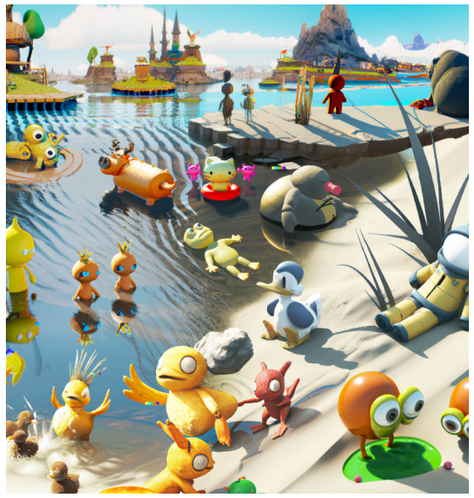
경쟁업자의 대응은 빨라진다

구글의 CEO인 선다 피차이는 챗GPT를 심각 단계의 코드레드(Code red) 수준으로 보고 구글 검색엔진에 미칠 수 있는 상황을 모니터링 중에 있다고 밝힌 바 있다. 또한 구글은 음악을 생성하는 AI 모델(Music LM)을 공개할 예정이었으나 취소했고, 그 구체적인 이유는 저작권 침해에서 자유로울 수 없다는 것으로 알려졌다. 다른 AI 모델에 공통되는 한계이자 문제인 데이터의 저작권 처리가 불분명하다는 점은 AI 모델의 강건성(Robustness)이 확보되지 않은 것으로 법률적인 분쟁이 예상되는 부문이다. 실제 미국에서 생성형 AI 모델에 대한 집단소송이 제기된 바 있다. 국내에서도 네이버 하이퍼클로바(Hyper Clover)라는 거대 AI 모델을 구축했으나, 다양한 법률적 이슈 등으로 인해 일반에 공개하지 않고 있다. 카카오나 LG 엑사온 등도 거대 AI 모델을 구축하고 있으나, 마찬가지로 외부 공개보다는 자체 사업 부문에 활용하고 있는 것으로 보인다. 다만, 챗GPT의 성공적인 런칭을 보면서 각 사는 공개적으로 운용할 것으로 예상되며, 법적 이슈를 포함한 사회적 이슈 대응은 더욱 커질 것으로 보인다. 실제 구글도 이에 대응한 모델을 공개한 바 있다.

챗GPT가 활용되고 있는 사례는 다양하다

무엇보다 놀라운 것은 텍스트형의 글을 작성하는 데 있다. 현재 공개된 챗GPT는 텍스트에 특화된 AI모델이지만 다양한 보고서, 계약서, 에세이, 시 등을 만들어내고 있다. 구글의 전문개발자 능력을 넘어서는 소스코드(Source Code)를 생성하고 있다. 원하는 스펙이나 알고리즘을 입력하면 그에 따른 코드를 생성한다. 또한 챗GPT는 아니지만 OpenAI의 또 다른 GPT인 Dall-E라는 이미지 생성 모델은 아래 그림처럼 이미지를 만들어낸다. 이처럼 GPT는 인간의 능력을 넘어서고 있다.

[그림 1] 달리2로 생성한 캐릭터



출처 : 달리2

참고로, 미드저니라는 이미지 생성 모델로 만든 그림이 지난해 미국의 미술전에서 수상을 한 사건이 있었으며, 이를 계기로 AI 모델이 만든 콘텐츠에 대해서는 표절이나 저작권 침해를 이유로 수상을 금지해야 한다는 여론이 일기도 했다. 이외에도 OpenAI의 GPT는 아니지만 동영상을 생성하는 AI 모델도 개발되고 있다. 영화나 광고에서 가장 중요한 것은 스크립트이며, 챗GPT가 스크립트를 만들어낸다는 점에서 영상 제작의 핵심적인 역할을 하는 것으로 볼 수 있다.

챗GPT의 저작권법적 이슈도 다양하다

챗GPT로 인한 고민거리는 저작권 문제를 어떻게 다뤄야 하는지이다. 인공지능이 스스로 만들거나 도구적으로 활용되는 경우 등 다양한 경우를 살펴볼 필요가 있다.

첫째, 챗GPT가 생성한 결과물은 저작권이 발생할까? 단정하기는 어렵다. 챗GPT는 사람이 아니다. 국내외 저작권법은 자연인인 사람만이 창작자라고 정의하고 있다. 우리 저작권법도 저작물을 '인간의 사상과 감정의 창작적 표현'으로 정의한다. 나투루 사건에서 원숭이의 셀카 사진에 미국 저작권법상 저작권이 없다고 판시한 것은 이를 반영한 결과이다. 따라서 챗GPT가 스스로 생성한 것이라면 저작물이 없다. 실제 챗GPT는 저작권이 있느냐는 질문에 자신에게 저작권이 없다고 답한다.

둘째, OpenAI나 챗GPT를 공동 저작자로 표기하는 것은 타당할까? 아니다. 우선 챗GPT는 사람이 아니므로 저작자가 될 수 없다. 표절에 대한 논란을 피하기 위한 출처표시 이상의 의미를 갖기 어렵다. 따라서 저작자로 표시할 것이 아니라 "챗GPT를 이용해서 제작했다"는 표시를 하는 것이 타당하다. 챗GPT 등을 저작자로 표시하는 것은 저작권법상 허위표시 금지에 위배될 수 있다. 일부 논문 등에서 공저자 형식으로 표시하는 것은 타당한 표현으로 보기 어려우며 사이언스(Science)지나 네이처(Nature)지 등에서는 챗GPT를 공동 저자로 기재한 논문을 승인하지 않고 있다.

셋째, 표절로 볼 수 있는 것은 아닌가? 맞다. 챗GPT를 이용했음에도 별도 표시 없이 자신의 성명만을 표시하는 경우에는 표절이 될 수 있다. 표절은 저작권이 있거나 없거나 상관없이 자신이 작성한 것이 아닌 글 등을 인용(Citation) 표시 없이 사용하는 것을 말한다. 따라서 저작자 표시가 아닌 "챗GPT를 이용해서 제작했다"는 표시를 해야 표절 시비에서 벗어날 수 있다.

저작권이 등록된 경우도 있다

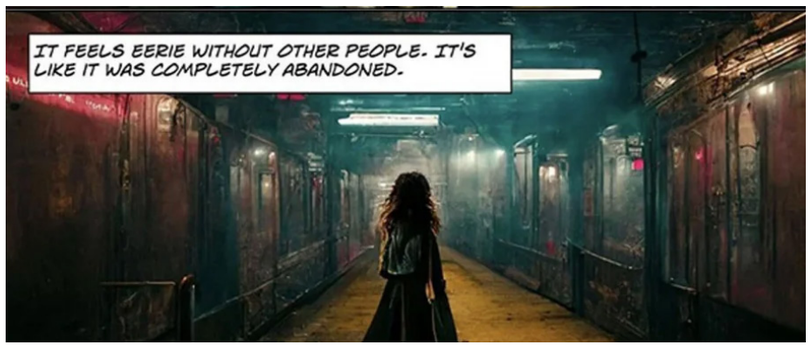
미국 저작권청에 이미지 생성 모델인 미드저니로 제작한 콘텐츠가 저작물로 등록된 경우도 있다. 저작권이 발생하기 위해서는 인간의 사상과 감정이 표현돼야 하며, 최소한의 창작적 기여를 의미하는 창작성이 있어야 한다. 미국 저작권청은 미드저니로 만든 소설인 『Zarya of the Dawn』를 저작물로 등록한 바 있으며, 미드저니를 이용하였더라도 사람이 가공한 것이기 때문에 인정된 것으로 볼 수 있다.

아래 그래픽 소설은 미국 저작권청에 등록됐는데, 2023년 현재 저작권청은 AI에 의해서 제작됐는지 등에 대한 소명을 요청했지만 저작권자가 소명하지 않아 등록이 취소될 것이라고 한다. 저작권청은 실질 심사가 아닌 형식적인 심사를 진행하지만 중요한 요건은 사람이 창작했는지 여부에 대한 확인이며 이러한 사실을 기재하지 않을 경우, 등록이 취소될 수 있기 때문이다.

프롬프트는 저작권법의 보호범위에 있는가?

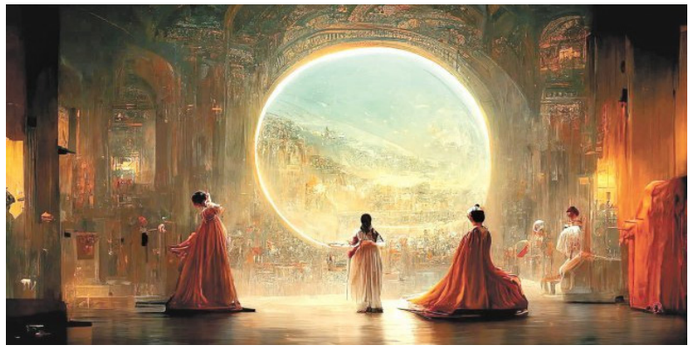
언어모델에 입력하는 명령어에 따라 콘텐츠 결과물의 질이 달라진다. 따라서 콘텐츠의 질을 높이기 위해 어떠한 프롬프트(Prompt)를 구성해야 하는지에 대해 고민하는 것이 프롬프트 엔지니어링(Prompt Engineering)이다. 프롬프트에는 수많은 노력이 들어간다. 프롬프트의 내용에 따라 결과물의 질(Quality)이 달라지기 때문이다. 또한 프롬프트에 수많은 명령어를 입력하고 수정하고 보완하는 과정에서 보다 완성도 높은 결과물이 만들어지기 때문이다. 이러한 이유 때문에 프롬프트 마켓이라는 별도의 시장이 형성되고 있다. 프롬프트 1개당 1~2달러에 이른다는 점에서 프롬프트가 갖는 가치는 적지 않다. 프롬프트 엔지니어링이라는 측면에서 이용자가 챗GPT의 화면에서 입력하는 명령어의 성격을 어떻게 볼 것인지도 중요한 논점이다. 프롬프트는 결과물과 어떤 관계를

[그림 2] 『Zarya of the Dawn』, 저작권 등록된 그래픽 소설



출처 : 『Zarya of the Dawn』

[그림 3] <스페이스 오페라극장>



출처 : Jason M. Allen with midjourney(2022)

가질 것인지도 고려해야 할 부문이다. 결합저작물인지, 챗GPT와 공동으로 창작한 결과물인지도 마찬가지다. 기본적으로 프롬프트가 다양한 내용을 만들어내는 조작이자 구성(Arragement)이며, 이는 인간의 사상과 감정이 담긴 표현임을 부인하기 어렵다.

위의 <스페이스 오페라극장>은 8시간 이상 소요돼 만들어낸 결과물이라는 점에서 작가의 창작적 기여는 충분하다고 생각된다. 결국, 창작적 기여와 그에 따른 결과물로 보건대 프롬프트의 저작물성은 충분하다.

편향, 환각효과에서 자유로울 수 있을까?

AI 모델은 공개된 데이터를 기반으로 학습하기 때문에 인간의 다양한 오류나, 문제점을 판단하지 않고 답습함으로써 편향적이고 차별적인 결과를 가져올 수 있다. 따라서 독성이 있거나 편향된 콘텐츠를 포함하지 않도록 이러한 모델을 교육하는 데 사용되는 초기 데이터를 신중하게 선택하는 것이 중요하다. 이외에도 부적절한 내용이 담겨있는 경우이다. 챗GPT는 내용 자체를 필터링하고 있지만, 모든 내용을 필터링하는 것은 아닌 것으로 보인다. 또한 OpenAI에서도 밝힌 바와 같이 결과물에 대해서 정확하지 않은 정보가 생성될 수 있다고 고지하고 있다. 챗GPT의 결과물에서도 사실과 다른 내용을 사실처럼 표현하고 있다는 점에서 환각효과가 문제 될 수 있음을 알아야 한다. 환각효과(Hallucination Effect)란 사람이 실제로 존재하지 않는 것을 보고, 듣고, 느끼고, 냄새 맡고, 맛보는 것과 같이 현실에 대한 거짓 또는 왜곡된 인식을 경험할 때 발생하는 현상이다. 챗GPT도 인터넷의 다양한 정보를 바탕으로 학습했고, 스스로 정보에 대한 진정성을 확인할 능력이 없기 때문에 언어적인 체계라는 면에서 답변은 잘하더라도 문맥과 의미에 대해서는 참과 거짓을 혼동할 수 있다. 이러한 면에서 볼 때 챗GPT도 환각효과(Hallucination Effect)에서 자유롭기는 어려울 것으로 보인다.

편향이나 연구윤리에 위배되지는 않는가?

앞서 살펴본 것처럼 윤리적으로 본인이 직접 작성한 것이 아닌 글을 자신의 이름으로 표시하는 것을 표절(Plagiarism)이라고 한다. 표절과 저작권법은 다른 이슈이며 타인의 저작물을 무단 이용하는 것은 저작권 침해 및 표절에 해당할 수 있으나, 저작권이 없는 글 등을 출처표시 없이 이용하는 것은 표절에 해당한다. 또한 챗GPT가 생성한 결과물을 인간이 자신의 것으로 표시하는 것은 표절 또는 연구윤리에 위배될 수 있으나, 저작권이 인정되기 어려운 상황에서 저작권 침해로 보기는 어려움이 있다. 참고로 학교 교육 현장에서 표절 여부를 판단하는 방법은 직접 구두로 테스트하거나, 직접 재현토록 하는 것이라고 할 수 있다.

또한 AI 모델은 공개된 데이터를 기반으로 학습하기 때문에 인간의 다양한 오류나 문제점을 판단하지 않고 답습함으로써 편향적이고 차별적인 결과를 가져올 수 있다. 따라서 독성이 있거나 편향된 콘텐츠를 포함하지 않도록 이러한 모델을 교육하는 데 사용되는 초기 데이터를 신중하게 선택하는 것이 중요하다.

문제는 대규모 언어모델이 갖는 한계에 있다

대규모 언어모델이 갖는 한계는 다양하겠지만 몇몇 항목에 대해서 정리해본다. 먼저 수많은 데이터가 사용된다. 둘째, 데이터에 담긴 문제가 반복되어 학습데이터로 사용될 수 있다. 셋째, 저작권이나 개인정보 처리가 되지 않은 결과물이 복제돼 현시(Display)될 수 있다. 다음으로, 차별이나 편향적인 결과와 표현의 자유가 충돌할 수 있다. 무엇보다, 데이터가 대규모라는 점이 가장 큰 한계이다. 다양한 데이터를 크롤링해 학습데이터로 만들기 때문에 수많은 문제가 내재된 데이터를 학습데이터로 활용할 수밖에 없다. 학습데이터의 정제과정에서 정제되지 못하는 편견이나 편향이라는 문제점은 그대로 AI 모델에 반영될 수밖에 없다. 데이터에는 다양한 계층, 시대, 분야의 문화적인 산물이 담겨있다. 그럼에도 불구하고 데이터의 정제과정에서 문화적, 세대 간, 계층 간 언어의 뉘앙스가 달라질 수 있다는 점을 간과한다. 그 결과 소설 등에서 여성에 대한 차별적인 표현, 인종에 대한 비난, 비윤리적인 행위의 정당화, 폭력을 넘어 살인을 미화하는 것, 동물 학대 등 다양한 내용이 학습데이터로 활용될 가능성을 배제하지 못한다. 학습데이터에 담긴 이러한 표현은 헌법상 표현의 자유 아래에서 보호받는 표현임을 부인하기 어렵다. 그렇지만, 해당 소설이나 작품 내에서 용인될 수 있는 표현이더라도, 독립되거나 맥락을 벗어나 인용하는 것은 의도성이 다르게 해석된다. 이러한 경우에 헌법적 가치를 어떻게 판단할 수 있을지는 의문이다. 인공지능에 의한 차별이라는 기본권을 해치는 것과, 인공지능에 의한 표현이라는 기본권을 지키는 것은 어디에 우선순위를 두어야 하는가? 이처럼 데이터에 담긴 문화적인 차별이 결과로써 재현된다는 점에서 과거의 데이터가 사용됐다고 항변될 수 있는 것은 아니다. 이미 새로운 콘텐츠를 만들기 위해 과거의 데이터를 사용했지만, 현재 상황에 맞게 표현되는 것이기 때문이다. 그렇지만 가짜뉴스와 같이 의도적인 왜곡이나 차별, 편향으로써 인간의 기본권에 대한 침해가 이루어지는 상황에서 표현의 자유를 지켜야 하는 것은 아니다.

향후 대응은 어떠해야 하나?

챗GPT 등에 대해서는 이제 논의가 되는 상황이다. 2016년 알파고가 AI의 확산에 기여했다면, 챗GPT는 우리의 일상생활에 직접적인 영향을 미친다는 점에서 문명사적 의의를 부여할 수 있다. 여러 가지 법률적인 이슈가 제기되지만, 기술규제가 아닌 기술현상을 지켜볼 필요가 있다. 따라서 다양한 논의를 통해 사회적 합의를 찾아가는 것이 필요하다.