

# 디지털 시대의 정보 탐색에 관한 기호학적 논의

## － 왜 우리는 ChatGPT에게 신뢰할 만한 정보를 기대하는가?

홍승혜\*

### 【 차 례 】

- I. 서론
- II. 디지털 시대의 정보 특성 및 정보 탐색의 변화
- III. 비판적 정보 수용에 영향을 미치는 요인들
- IV. 대화형 인공지능 언어 모델과 정보 탐색
- V. 결론

### 국문초록

본 연구는 디지털 시대의 정보 탐색과 관련하여 크게 두 가지 질문에 기반하여 논의를 전개한다. 첫째, 비판적 정보 수용은 현대인에게 끊임없이 요구되는 디지털 리터러시 역량 중 하나인데, 이를 개발하는 데 여전히 어려움을 겪는 이유는 무엇인지에 관한 물음이다. 이와 관련하여 본 연구는 특히 사회·문화 시스템이 만들어낸 텍스트에 대한 습관화된 믿음과 정보 공유의 주요 매개체인 언어 및 문자와 같은 상징기호의 속성에 주목하여 논의한다. 둘째, 디지털 시대의 또 다른 정보 탐색 도구로 등장한 ChatGPT와 같은 인공지능 언어 모델에게 왜 우리는 신뢰할 만한 정보를 기대하는가에 대한 물음이다. 이에 대해 본 연구는 ChatGPT는 웹 검색과 달리 대화형 모듈이라는 점에서 대화의 격률 준수가 요구되며, 인공지능이 정확한 정보를 제공해야 한다는 기대의 기저에는 모델의 언어적 유창성이 있음을 논의한다. 더불어 상징기호에 기반해 작동하는 인공지능의 메커니즘 특성상 불확실한 정보가 생성될 가능성은 불가피하며, 이를 고려하여 인공지능을 활용한 정보 탐색을 탐구의 지표로서 활용하는 방향을 제안한다. 이 연구는 언어와 문자라는 상징기호를 통해 이루어지는 현대인의 정보 탐색

\* 단독저자, 고려대학교 강사, [happy\\_sh@korea.ac.kr](mailto:happy_sh@korea.ac.kr)

습관을 기호학적 관점에서 재고하고, 인공지능 언어 모델 개발은 결국 인간과의 소통을 목적으로 하므로 모델의 개발 및 활용에 있어 언어기호에 대한 본질적 이해가 선제되어야 함을 보인다는 데에서 의의를 찾고자 한다.

열쇠어 : 챗GPT, 대화, 대화의 격률, 디지털 리터러시, 정보, 지표, 상징기호, 아비투스

## I. 서론

2022년 11월 30일 ChatGPT가 처음 공개되고 난 직후 모델의 뛰어난 언어 생성 능력은 산업계뿐만 아니라 일반 대중들의 이목을 집중시켰다. 그 와중에 ChatGPT가 실재하지 않는 사실이나 현상에 대해서도 그럴듯한 설명을 산출한다는 치명적 문제점이 제기되었다. 언론을 통해 가장 많이 회자되었던 대표적인 사례로 ‘세종대왕 맥북 투척 사건’에 대한 생성 결과를 들 수 있다. 한국인 성인이라면 그것이 존재할 수 없는 사건임을 단번에 알 수 있지만, ChatGPT는 아주 그럴듯하게 해당 사건에 대해 설명해낸 것이다. 실제 존재하지 않는 것을 마치 존재하는 것처럼 말을 한다는 측면에서 이는 ‘환각(hallucination)’ 현상으로 일컬어지고 있다.

다만 여기서 주목할 점은, 해당 사례가 ChatGPT의 치명적 단점이자 환각과 다름없는 것으로 치부된 이유는 ‘세종대왕 맥북 투척 사건’에 대한 생성 결과를 대중들은 ‘이야기(story)’가 아닌 ‘정보(information)’로 수용했기 때문이라는 점에 있다. 만약 ChatGPT에게 기대한 것이 ‘세종대왕 맥북 투척 사건’을 주제로 하는 상상 속 가상의 이야기 생성이었다면 그 결과물을 무척 훌륭하다고 평가했을 것이다. 그러나 대중들은 인공지능 언어 모델의 유창한 언어 능력 그 자체보다 모델이 생성한 내용의 정확성에 더욱 주목하였다. 다시 말해, 대중들은 뛰어난 언어 생성 능력으로부터 신뢰할 수 있는 정보 제공 능력을 전제하고 있는 것이다.

한편 인터넷에 신뢰할 수 없는 정보가 많다는 사실은 이미 잘 알려져

있다. 이러한 문제점을 인식하고 있음에도 인터넷의 생리적 특성상 실시간적인 데이터의 생성 속도에 맞추어 이를 선별하고 통제할 재간은 그 누구에게도 없다. 따라서 정보를 비판적으로 선별하는 것은 오롯이 수용자의 몫이 되고 있다. 그러나 우리는 여전히 인터넷 검색을 통해 접하는 정보를 비판적으로 받아들이는 데 취약하다. 존재하지 않는 사실이나 잘못된 내용을 담은 익명의 진술에 쉽게 속는 것이다. 이 와중에 인공지능이라는 존재까지 정보 제공의 장으로 들어섰다. 이러한 측면에서 디지털 시대의 정보 탐색에 관한 재고가 요구된다.

이에 본 연구는 크게 두 가지 질문에 기반하여 논의를 전개하고자 한다. 첫째는 디지털 환경에서 비판적 정보 수용은 왜 어려운 것인지 그 근본적 원인을 탐구하고자 하는 질문이다. 둘째는 우리는 왜 대화형 인공지능 언어 모델에게 신뢰할만한 정보의 제공을 기대하는 것인가이다. 본격적 논의에 앞서 2장에서는 정보의 개념과 현대인의 정보 탐색 양상을 간략히 살핀다. 3장에서 기존의 정보 탐색 습관과 정보의 주요 매개체인 상징기호의 속성을 분석함으로써 첫 번째 질문의 답을 찾는다. 4장에서는 대화형 인공지능 언어 모델의 특성과 메커니즘에 대한 이해를 토대로 두 번째 질문의 답을 찾고, 비판적 정보 수용에 대한 긴장성을 낮춰 인공지능을 정보 탐색에 활용할 수 있는 방안을 제안한다. 5장은 결론으로 논의를 요약하고 본 연구의 의의를 밝힌다.

## Ⅱ. 디지털 시대의 정보 특성 및 정보 탐색의 변화

### 1. 형식으로서의 정보

디지털 리터러시의 핵심 역량 중 하나로 디지털 매체를 통한 정보의 비판적 수용을 들 수 있다.<sup>1)</sup> 이는 인터넷이 일상화되면서부터 현대인에

---

1) 김정희, 김광재, 이숙정, 「모바일 환경에서의 미디어 리터러시 구성 요소와 세대 간

게 끊임없이 요구되는 역량이다. 정치, 종교, 성별의 측면에서 편향된 정보 혹은 오류가 있거나 검증되지 않은 정보들이 혼재되어 있을 수 있으므로 이를 선별적으로 취할 수 있어야 한다는 것이다. 이와 관련하여 아이어톤&포세티(2020)에서는 공익성을 해치는 정보의 유형을 크게 ‘잘못된 정보’, ‘허위 정보’, ‘유해 정보’로 구분한다.

- 잘못된 정보: 사실은 아니지만 유포하는 사람은 진실이라고 믿는 정보
- 허위 정보: 허위일 뿐 아니라 그것을 유포하는 사람도 허위란 사실을 알고 있는 정보
- 유해 정보: 실제로 바탕을 두고 있지만 사람, 조직, 혹은 국가에 해를 끼치기 위해 사용되는 정보<sup>2)</sup>

‘잘못된 정보’와 ‘허위 정보’는 그 정보가 전달하는 내용이 사실이 아니라는 점에서 ‘유해 정보’와 구분된다. ‘유해 정보’는 비록 그 내용이 사실일지언정 그 정보가 수용자에게 부정적 영향을 미친다는 점에서 경계되어야 하는 정보이다. 한편 ‘잘못된 정보’와 ‘허위 정보’는 이를 유포하는 사람이 그 정보의 사실 여부를 인식하느냐에 따라 구분된다. 여기서 주목할만한 점은 그 내용이 사실이 아닌 경우에도 ‘정보’라 일컬어지고 있다는 것이다. 이러한 측면에서 ‘정보’의 범주화는 내용적 특성보다 특정한 ‘형식(form)’에 기반함을 가늠할 수 있다.<sup>3)</sup> 따라서 ‘정보의 비판적 수용’이라 함은 정보의 형식에 좌우되지 않고 수용할만한 내용을 선별하여 취할 수 있음을 의미한다고 볼 수 있다.

그렇다면 정보의 형식은 무엇인가? 퍼스의 설명에 따르면, 상징(symbol)은 그 대상에 대해 삼중의 참조를 갖는다. 첫째는 그것이 재현

미디어 리터러시 격차」, 한국방송학보33(4), 2019, 5~36쪽.

2) 체릴린 아이어톤 & 줄리 포세티, 『저널리즘, 가짜뉴스 & 허위정보 : 저널리즘 교육과 훈련을 위한 핸드북』, 김익현 역, 서울: 한국언론진흥재단, 2020, 77쪽.

3) 여기서 ‘형식’이란 어떤 대상의 사실 여부와 관련 없이 모든 경우에 무언가를 그렇게 존재하도록 하는 자질들 및 그것의 존재 방식을 의미한다(cf. W 1:307).

하는 실제 대상 자체, 둘째는 해당 대상의 공통적 특성, 셋째는 그것의 해석체 또는 대상에 대해 알려진 모든 사실이다(W 2:82-83). 첫째는 정보적 넓이로 외연과 같으며, 둘째는 정보적 깊이, 셋째는 ‘그 상징이 주어 또는 서술어인 명제(synthetical propositions)의 합 또는 그 상징에 관한 정보’에 해당한다(ibid.). 이때 상징의 세 번째 참조 유형을 정보의 형식으로 참조할 수 있다. 즉, 특정 개념에 대한 질의의 답이 그 개념어를 주어나 서술어로 갖는 명제의 형태로 재현된다면 그 형식의 재현 자체가 정보로 인식될 수 있는 것이다.

그러나 실제 우리가 정보로서 기대하는 것은 단순히 형식이 아니다. 이와 관련하여 ‘정보’의 사전적 정의를 살펴보자. <표준국어대사전>에 따르면 정보는 “관찰이나 측정을 통하여 수집한 자료를 실제 문제에 도움이 될 수 있도록 정리한 지식 또는 그 자료”를 의미한다. 이때 ‘실제 문제에 도움이 될 수 있도록 정리된 지식’이라는 기술은 퍼스의 실용주의(pragmaticism)를 상기시킨다. 퍼스의 실용주의는 경험할 수 있는 효과에 관한 것이다(CP 5.438). 즉, 한 사람에게 제공된 정보가 그 사람이 행동하거나 사고하는 데 있어 어떠한 실제적 효과를 주지 않는다면 그것은 정보라고 말하기 어렵다. 퍼스는 또한 정보의 특성에 관해 다음과 같이 설명한다.

“만약 당신이 나에게 어떤 참인 것(truth)에 대해 말하였는데 내가 이미 알고 있는 것이라면 그것은 정보가 아니다. 만약 그것이 내가 믿을 이유가 없는 것이라면, 당신은 내가 관심을 갖지 않는 우주에 대해 말하는 것이며, 또한 당신이 말하는 것은 나에게 아무런 의미가 없다(MS [R] 463:13).”

이처럼 정보로서 수용되는 것은 세상의 모든 참인 명제가 아니라 한 개인에게 영향을 줄 수 있는 명제에 해당한다. 즉 어떤 대상에 관한 사실을 알게 됨으로써 개인의 행동이나 사고를 변화시킬 수 있는 것이 정

보의 속성이라고 할 수 있다. 그러나 정보로서 정리된 내용이 자신이 직접 관찰하여 도출한 명제가 아니라면, 그 점이 비판적 정보 수용의 취학점으로 작동할 수 있다. 잘못된 정보나 허위 정보일지라도 수용자는 이를 의심할 기반이 없어 그것을 그대로 정보로서 수용하는 경우가 발생하는 것이다. 따라서 무언가를 정보로 수용하기 위해서는 대상을 직접 탐구하거나 그것을 정보로 수용할 만한 근거가 있어야 한다.

이처럼 한 개인이 정보로 수용하는 것에는 참이라는 전제에서 기인하는 신뢰성과 실용성의 자질이 부착되어 있다. 그 자질의 충족은 정보 그 자체가 아닌, 관찰과 측정을 통해 실제 문제에 도움이 될만한 것으로서 내용을 정리한 최초의 정보 유포자로부터 기인할 것이다.<sup>4)</sup> 정보 유포자에 대한 신뢰성이 정보에 대한 신뢰성으로 연결되는 것이다. 따라서 수용자는 신뢰할 수 있는 출처로부터 정보를 찾을 것이나, 그 출처를 신뢰하기 어려운 경우 검증된 정보를 참조하여 새로운 정보를 검증해야 할 필요가 있다. 이때 정보를 매개하는 기호의 재현적 특성 및 정보가 교환되는 맥락에 대한 수용자의 이해도에 따라 그 결과가 달라질 수 있다.

예를 들어, 개별 포장된 사과 제품에 “이 사과는 세척 과정을 거쳐 위생적으로 포장되었다.”라는 안내문이 제시되어 있을 때, 그 사과를 바로 먹거나 세척을 한번 거친 뒤 먹을 수도 있다. 안내문을 믿고 세척 없이 사과를 먹는다는 것은 해당 사과를 생산한 업체를 신뢰한다거나 혹은 다른 소비자들도 의심 없이 그 제품을 취하기 때문일 수도 있다. 위생 시설로 검증되었다는 HACCP과 같은 마크는 그러한 신뢰의 강력한 기반이 될 것이다. 그러나 만약 그 업체에 대해 충분히 알려진 바가 없다면, 또 이를 뒷받침할 수 있는 근거가 없다면, 과연 그 사과는 세척하지 않고도 먹을 만한 것인지 그 업체의 생성 공정에 대해 추가적인 검증을 하

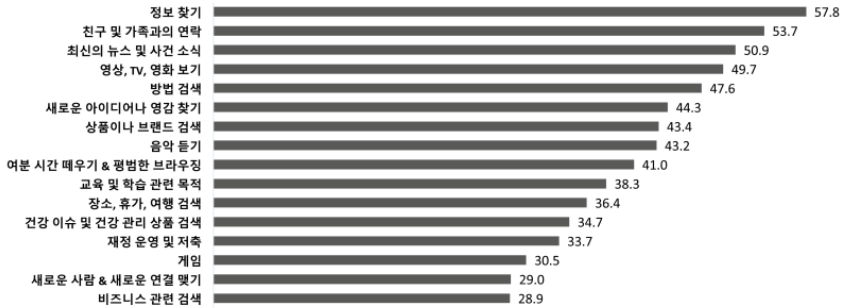
---

4) 물론 현재의 정보가 미래에는 잘못된 정보가 될 수도 있다. 또 과거에는 잘못된 정보였지만, 현재에는 참일 수도 있다. 이처럼 정보는 잠재적으로 참일 수도 또 거짓일 수도 있다는 ‘불확실성’을 갖는다.

거나 직접 닭아 먹는 방법을 취해야 할 것이다.

## 2. 디지털 환경에서의 정보 탐색

그렇다면 디지털 환경은 정보의 탐색에 어떤 변화를 가져왔는가? 2023년 “We are Social”의 연례 보고에 따르면 전 세계 인터넷 사용자는 최초로 그 수치가 집계되었던 2013년 이래 지속적으로 증가하는 추세에 있다. 더불어 같은 보고서에 제시된 [그림 1]을 살펴보면, 사람들이 인터넷을 사용하는 목적 가운데 ‘정보 찾기(finding information)’가 가장 높은 비중을 차지함을 알 수 있다. GWI 연구팀은 올해 보고서에서 인터넷 사용자의 57.8%가 정보를 찾을 때 온라인 자원을 활용한다고 발표하였다.<sup>5)</sup> 해당 통계를 통해 인터넷은 현대 사회에서 주요한 ‘정보의 보고’로 인식되고 있음을 알 수 있다.



[그림 1] 인터넷 사용 목적 (We are Social, 2023)

인터넷을 “정보를 얻는 중요한 수단”으로 인식하는 것은 이미 2000년 도부터 만연했다. 이제는 모르는 것이 있으면 가장 먼저 인터넷을 통해 검색해보는 것이 당연한 절차가 되었다. 심지어 검색을 해보지 않고 그

5) We are Social, *THE CHANGING WORLD OF DIGITAL IN 2023*, 2023.

것에 대해 잘 알고 있는 누군가에게 직접 묻는 행위에 대해 부정적 프레임을 씌우는, ‘핑거 프린스/프린세스(finger prince/princess)’라는 신조어도 탄생했다. 검색 한 번이면 알 수 있음에도 손가락 몇 번 움직이기 귀찮아 다른 사람에게 물어본다는 의미를 내포하는 표현이다. 이 역시 우리가 정보 획득에 있어 인터넷 검색에 얼마나 의존하고 있는지, 또 인터넷 정보를 얼마나 신뢰하고 있는지를 단적으로 보여주는 사례라 할 수 있다.

물론 그 가운데 이미 참으로 검증된 사실과 역사적으로 합의에 이르러 타당성을 갖는 정보들이 상당 부분을 차지한다. 특정 대상에 대해 알고 싶을 때 인터넷을 검색해보면 그 대상을 설명하고 있는 텍스트를 찾을 수 있다는 점에서 인터넷은 ‘정보’의 장으로 인식되고 있다. 그러나 불특정 다수에 의해 작성되어 수십 년간 인터넷에 누적되어온 데이터는 세상의 모든 참을 대변하지 않는다. 이러한 측면에서 현대 사회의 정보 탐색은 검색 결과가 신뢰 가능한 것인지에 대한 고려 없이 일단 알고 싶은 그 대상에 대한 이해의 깊이가 확장된다는 데에 초점이 맞춰져 있다고 할 수 있다.

한편 ChatGPT와 같은 생성형 인공지능 언어 모델이 일상생활 속에 점점 더 깊숙이 들어오며 따라 정보 탐색 양상도 변화할 것임을 어렵지 않게 예측할 수 있다. 생성형 인공지능 언어 모델이 등장하기 이전에 검색 엔진을 통해 정확한 정보를 얻기 위해서는 몇 차례에 걸쳐 탐색 과정을 진행해야 했다면, 이제는 그러한 과정이 한 차례의 질의응답으로 단축될 수 있기 때문이다. 흥미롭게도 대중들은 인공지능이 제공하는 정보의 정확성에 있어 인터넷 검색과 달리 엄격한 평가 잣대를 들이밀고 있다. 그러나 만약 인공지능 언어 모델에게 자신이 전혀 모르는 것에 대해 묻는다면, 생성된 결과를 비판적으로 수용하지 못하는 것은 마찬가지다.

이때 주목할 부분은 인공지능 언어 모델을 포함해 온라인에서 획득하는 대부분의 정보는 ‘문자’를 매개한다는 점이다. 물론 최근 이미지와 영



상 정보 역시 많은 영향력을 미치고 있지만, 디지털 환경에서 정보의 교류는 여전히 문자에 의존적이다. SNS와 유튜브에서 공유되는 이미지나 영상 역시 텍스트에 기반하여 검색되며, 그에 관한 설명도 텍스트로 전달된다. 인터넷은 통시적 차원에서뿐만 아니라 공시적 차원에서의 소통 범위와 가능성을 더욱 광범위하게 확장하는 데 그 혁신적 기능이 있다. 그러나 시공간을 망라한 소통의 가능성은 본래 문자의 속성에 기반한다는 점에서 온라인에서 텍스트 정보의 영향력은 더욱 강력해진다.

### Ⅲ. 비판적 정보 수용에 영향을 미치는 요인들

#### 1. 텍스트 아비투스

인간은 활자 혁명 이래 개인의 경험 세계를 넘어서는 세상에 관한 다양한 정보를 주로 책을 통해 습득해왔다. 글을 읽는 법을 배우기 전에는 어른의 말에 의존하여 정보를 얻지만, 이후 책으로 학습 수단이 전이되는 것이다. 동서고금을 막론하고 모든 교육의 장에서 ‘독서’는 핵심 키워드이다. 이러한 관습은 책이 담고 있는 텍스트에 대한 신뢰를 전제한다. 초·중·고 전 과정에서 우리는 교과서, 참고서, 유인물 등 텍스트 의존적인 학습을 이행한다. 이러한 텍스트 자원을 쉽게 의심하지 못하며, 시험 문제의 정답은 텍스트로 기록된 것의 유무로 판별된다. 이러한 과정에서 텍스트에 대한 막연한 신뢰는 내재화될 수밖에 없다.

성인이 된 이후에도 신문기사나 전문서적을 탐독하는 것과 같이 문자에 의존한 학습 및 정보의 습득이 지속된다. 따라서 언론 기관 및 출판사는 저마다의 신뢰성을 확보하고 지키기 위한 노력을 가한다. 오보나 검증되지 않은 저자로 인한 논란은 이윤 창출과 직결되는 브랜드의 명성 및 운영에 치명적으로 작용하기 때문이다. 따라서 그들은 자신들이 출판하는 콘텐츠에 대한 자체적인 검증 시스템을 운영한다. 논문의 경우 주

로 각 분야에 학자와 전문가가 기호 생성의 주체가 되며, 엄격한 심사 및 동료평가를 통해 타당성을 검증하는 과정을 거친다. 이러한 자체적인 검증 절차를 통해 각 매체는 자신들이 매개하는 정보의 신뢰성을 뒷받침한다.

이처럼 텍스트에 대한 습관화된 신뢰의 기저에는 서적, 신문, 방송과 같은 미디어의 영향력이 있다. 이에 기반하여 텍스트의 신뢰성에 대한 잠재된 믿음이 유지되는 것이다. 일례로 우리는 신문기사가 실제 벌어진 사건이나 사실을 기술하고 있음을 전제하고, 그 내용을 의심하지 않고 수용한다. 신문이라는 미디어를 신뢰하는 것이다. 신문이 갖는 신뢰성을 이용하는 광고 전략도 있다. 신문기사의 형식을 온전히 갖춘 채 일간지에 실려있지만, 그 내용을 살펴보면 특정 상품이나 장소에 대한 홍보임을 포착한 경험은 누구나 있을 것이다. 이러한 기사형 광고는 상업성보다 공익성을 띤다는 착각을 불러일으킨다. 신문기사의 형식을 모방한 것만으로 신뢰할 수 있는 정보로 인식되기 쉬운 것이다. 일종의 프레이밍 효과이다.

정보를 제공해주는 미디어가 신뢰할만한 시스템과 권위를 갖추고 있는 경우, 수용자는 비판적 정보 수용에 더욱 취약해진다. 미디어의 공신력이라는 믿음에 기반하여 우리는 각 미디어를 통해 전달되는 텍스트가 재현하는 것에 대해서도 막연한 믿음을 갖게 되는 것이다. 퍼스는 믿음을 고정하는 네 가지 방법으로, 고집의 방법, 권위의 방법, 선형적 방법, 과학의 방법을 제안한다(CP 5.377-5.385). 이 가운데 미디어를 통해 얻는 정보에 대한 신뢰는 두 번째 방법인 권위에 의한 믿음으로 이해할 수 있다. 이 경우 믿음의 기반은 정보의 수용자가 아닌 신뢰할만한 것으로 인식된 미디어라는 외부에 있게 된다. 이러한 측면에서 권위의 방법은 진정한 탐구의 방법이 되기 어렵다.<sup>6)</sup>

사회·문화적 시스템을 기저로 하는 텍스트에 대한 무의식적 신뢰는

---

6) 코르넬리스 드발, 『퍼스 철학의 이해[개정판]』, 이윤희 역, HUIINE, 2019. 160쪽.

우리의 관점, 행동, 사고방식을 좌우하는 ‘아비투스(habitus)’의 한 측면에 해당한다. 해당 개념을 주창한 피에르 브루디외(Pierre Bourdieu)의 진술에 따르면, 아비투스는 “사회화된 주관성(a socialized subjectivity)”이자 “체현된 사회성(the social embodied)”이다.<sup>7)</sup> 개인이 성장해온 사회와 문화의 기질이 몸과 의식에 자연스럽게 흡수되어 개인이 행동하고 사고하는 방식을 가이드하는 것이다. 성장 과정에서 세상에 대한 지식을 습득하는 주요 통로는 교육자가 집필한 교과서와 콘텐츠가 검증된 책이었기에, 그러한 습관으로부터 축적된 텍스트에 대한 막연한 신뢰는 텍스트로 기록된 것들에 의심을 가하기보다 의심하지 않고 수용하도록 가이드할 가능성이 높다.

텍스트에 대한 신뢰를 바탕으로 정보를 획득해왔던 아비투스에 의해 우리는 변화된 사회의 장에서 정보를 수용하는 태도를 쉽게 변화시키지 못하고 있다. 아비투스는 오랜 기간 천천히, 또 무의식적으로 체현된 기질이라는 측면에서 우리가 거주하는 사회적 환경이 변화하여도 쉽게 사라지지 않기 때문이다.<sup>8)</sup> 하지만 브루디외가 강조하길 아비투스는 단순히 일반적인 관습이나 습관이 아닌 “기저의 원칙(underlying principle)”으로서 관습을 ‘생성(generate)’한다.<sup>9)</sup> 이러한 측면에서 현대인이 정보를 습득하는 경로가 디지털 플랫폼, 인공지능 언어 모델로 확장됨에 따라 텍스트에 대한 믿음은 재고되어야 마땅하다. 디지털 환경에서 기호 생성자 및 정보의 출처에 대한 표시는 빈번히 생략되며, 그 정보에 대한 검증 절차도 필수적으로 요구되지 않기 때문이다.

그렇다면 왜 현대인은 새로운 변화의 흐름에 따르는 관습의 생성에 어려움을 겪는가? 현재 수많은 디지털 플랫폼을 통해 공유되는 글과 인공

7) P. Bourdieu & L. Wacquant, *An Invitation to Reflexive Sociology*, L. Wacquant (trans.), Cambridge: Polity, 1992. pp.127~128.

8) Michael James Grenfell, *Pierre Bourdieu: Key Concepts: Vol. 2nd ed.* Routledge, 2014, p.58.

9) *ibid.*, 55.

지능 언어 모델이 생성하는 텍스트는 신뢰할만한, 즉 공신력있는 뉴스 혹은 신문기사의 형식을 취하고 있지 않음에도 정보로서 수용되고 있다. 이는 우리가 관습을 변화시키지 못했다가보다는 ‘인터넷’이라는 새로운 정보 교환 양식에 의해 가이드된 결과라고 할 수 있다. 인터넷을 통해 획득할 수 있는 정보의 다양성은 특출난 한 개인의 지식 세계에 견줄 수 없을 만큼 광범위하다. 검색을 통해 관련된 정보를 병렬적으로 찾아낼 수 있는 시스템은 그 자체로 현대 사회의 엄청난 매체 권력에 해당한다. 따라서 이러한 시스템이 일상화됨에 따라 행동하고 사고하는 방식이 습관화된 것이다.

한편 ChatGPT와 같은 대화형 인공지능 언어 모델은 기존의 인터넷 검색과 달리 또 다른 정보 교환 양식의 영향을 받게 되는데, 이에 관해서는 4장에서 논의하도록 하겠다.

## 2. 정보의 매개체로서 상징기호의 속성

정보가 전달되는 신문이나 책과 같은 매체가 아닌 그 매체의 콘텐츠를 구성하는 상징기호가 갖는 특성이 정보의 신뢰성에 어떤 영향을 미치는지에 대해 생각해볼 수 있다. 앞서 논의한 사회·문화적인 차원에서 형성된 텍스트에 대한 신뢰에 선행하여, 기호 자체의 특성이 정보를 비판적으로 수용하는 데 미칠 영향력을 고려할 수 있기 때문이다. 인터넷의 정보는 주로 문자로 공유되며, 영상이라 하더라도 음성 및 음성을 재현하는 자막이 정보 전달에 있어 주요한 역할을 수행한다. 이미지 영상만으로는 정보로서 기능하는 데 제약이 있다. 그렇다면 언어와 문자 같은 상징기호가 갖는 해석적 효과가 해당 기호들의 구성체를 통해 재현되는 대상에 대한 신뢰성에 어떠한 영향을 미치는지 논의해보자.

우리가 문자로 기록된 것을 읽는다는 것은 결국 “음독이든 묵독이든 텍스트를 음성으로 옮기는 일”이다.<sup>10)</sup> 문자를 통해 음성 이미지가 우선

적으로 상기됨으로써 우리는 기호와 대상간 습관화된 법칙에 기반하여 그 기호가 재현하는 대상, 즉 텍스트가 담고 있는 내용을 의식 가운데 전개한다.<sup>11)</sup> 현대인에게 문자와 음성은 강하게 동기화되어 있어, 문자로부터 음성을 상기하여 그로부터 대상을 떠올리는 것은 기호가 해석자에게 미치는 효과로서 ‘직접적 해석체(immediate interpretant)’에 해당한다. 직접적 해석체는 “분석 이전의 최초의 그것 전체로서, 직접적으로 유의미한 가능한 효과”(MS 339d: 546)이다.<sup>12)</sup> 의심을 가하기 전에 이미 의식 속에 명제가 펼쳐지는 것이다.

이러한 ‘직접적 해석체’는 ‘역동적 해석체(dynamic interpretant)’로 나아간다. 역동적 해석체는 “모든 정신이 실제로 기호에 대해 행하는 모든 (형태의) 해석”(CP 8.315)이다.<sup>13)</sup> 따라서 모든 직접적 해석체가 온전히 역동적 해석체로 수용되지 않을 수 있으며, 그것을 개인의 내면세계에 유의미한 명제로 수용할지는 해석자에 달려있다. 즉각 효과를 미치는 직접적 해석체와 달리 정신의 해석인 것이다. 만약 발화자를 믿을 수 없다거나, 직접적 해석체가 자신의 믿음 세계와 완전히 배치된다면 직접적 해석체가 그대로 역동적 해석체로 나아가지 않을 것이다. 만약 어떤 의심이 생겨나지 않는다면 그것은 역동적 해석체의 하위 분야인 ‘논리적 해석체(logical interpretant)’로 나아가, 정보의 확장에 기여할 것이다.

논리적 해석체는 “기호의 궁극적 효과”(MS 339d:547)로서 한 단계 더

10) Walter J. Ong, *Orality and literacy: the technologizing of the word*, London; New York: Routledge, 2002, p.35.

11) 퍼스는 해석체의 도움에 의해서만 실현될 수 있는 특성을 갖는 기호인 상징을 ‘진정한 기호(genuin sign)’라 칭한다(CP 2.92). 상징은 해석체가 제거되면 그것이 기호로 작동하게 하는 특성을 잃게 된다(CP 2.304). 상징의 예로는 모든 단어, 문장, 책, 관습적 기호 등이 있다(EP 2:274).

12) 퍼스의 또 다른 해석체 유형에 따르면 직접적 해석체와 같은 해석체 삼분법의 제일요소로 ‘즉각적(felt) 해석체’(CP 8.369-372)와 ‘습관적(naive) 해석체’(MS 499:47)가 있다. 이에 대해 리슈카는 즉각적 해석체는 일례로 어떤 것이 즉시 지각되는 것, 습관적 해석체는 인지나 분석 없이 즉각적으로 지각되는 것이라고 해제한다(리슈카, 2019:277).

13) 제임스 야콥 리슈카, 『퍼스 기호학의 이해[개정판]』, 이윤희 역, HU:iNE, 2019, 79쪽.

나아가 ‘최종적 해석체(final interpretant)’에 도달할 수 있다.<sup>14)</sup> 한편 기존에 의심 없이 정보로서 받아들였던, 텍스트로 기록된 누군가의 진술이 사실이 아닌 것으로 밝혀지거나 허황된 것으로 증명된다면, 해석자는 그 기호 생성자를 더는 신뢰하지 않고 그가 생성하는 기호가 재현하는 대상에 의심을 가할 것이다. 즉, 그 사람이 생성하는 기호로부터 발생하는 직접적 해석체는 최소한 개인의 내면세계에서 역동적 해석체로 나아가지 못하는 것이다. 그러나 기호 생성자가 누구인지 알지 못한 채, 매체가 주는 권위성에 기대어 있는 상태라면 우리는 습관적으로 기호가 재현하는 대상을 믿게 된다. 직접적 해석체가 곧장 역동적 해석체로 수용되는 것이다.

이처럼 상징기호를 통해 정보를 얻는 경우,<sup>15)</sup> 대상과의 인접성을 전제하는 지표기호와 달리 대상의 실존성 혹은 사실성을 의심하기 어렵다. 감각적으로 경험 가능한 세계에 대한 지각적 판단은 ‘지각편린(percept)’에 기반하므로 사실성이 전제된다. 따라서 관찰에 기반한 언어적 해석이 곧 정보가 된다. 하지만 의식세계에서 이루어지는 상징기호에 대한 해석은 일차적으로 습관화된 법칙에 따라 기호가 재현하는 것을 상기하는 것이다. 그 대상은 언어를 매개해서만 알 수 있다. 의식 속에 불러 들어오는 명제 형태의 직접적 해석체는 이미 정보의 형식을 취하고 있어, 대상을 직접 관찰하거나 또 다른 언어기호가 해석에 개입하지 않는 이상, 기호가 재현하는 대상에 의심을 가하기 어렵다.

이때 정보 공유와 관련된 또 다른 맥락적 요소가 해석자가 정보를 수용하는 데에 동시에 영향을 미친다. 앞서 논의하였듯이 정보가 전달되는 매체 혹은 정보를 제공한 사람에 대한 신뢰가 정보 수용에 관여하는 것

14) 해석체 분류는 학자에 따라 관점에 따라 상이한 측면이 있다. 필자는 ‘감정적, 활력적, 논리적 해석체’를 역동적 해석체의 하위 유형으로 구분하는 드발(2019)의 의견에 따르고자 한다. 그러나 그 하위 부류 가운데에서도 감정적 해석체는 직접적 해석체에, 논리적 해석체는 최종적 해석체에 가까이 위치한다는 측면에서 리슈카(2019)의 구분과 완전히 배치되지는 않는다고 본다.

15) 단어들의 조합들 또한 상징에 해당한다(W 1:468).

이다. 충분히 신뢰하는 출처는 의심이나 검증의 필요성을 약화시킨다. 반대로 정보가 공유되는 맥락이 비사실성을 전제하는 경우 그럴듯한 진술도 정보로 인식되지 않는다. 대표적인 사례로 소설을 들 수 있다. 소설이라는 장르는 사실성을 기대하지 않는다. 오히려 소설이라는 장르 안에서 재현되는 모든 것들은 가상성을 전제한다. 따라서 소설 텍스트상에서 기술되는 바가 학습의 정보로는 활용되거나 그렇게 되기를 기대하지 않는다.

이러한 측면에서 ChatGPT가 만들어내는 ‘세종대왕 맥북 투척 사건’과 같은 비현실적인 이야기에 비판적으로 접근할 필요성이 제기되지 않는다. 생성 결과를 단순히 가상성을 포함한 이야기로 받아들일 수 있다면 말이다. 이러한 관점에 따르면 인공지능 모델을 통해 정보를 탐색하고 수용하는 자세도 달라질 것이다. 즉 인공지능 언어 모델의 성능 평가는 수용자가 그 생성 결과를 어떠한 맥락과 관점에서 받아들일지가 관건이 된다. 다만 현재 대중에게 인공지능 언어 모델은 ‘소설가’보다는 ‘교과서’와 같은 존재로 인식되는 것으로 보인다. 그렇다면 인공지능 언어 모델이 신뢰할만한 정보를 제공할 것이라는 혹은 그래야 한다는 기대는 어디에서 오는가?

## IV. 대화형 인공지능 언어 모델과 정보 탐색

### 1. ChatGPT에 대한 대중의 인식과 언어적 유창성

기존의 인터넷 검색 엔진과 달리 ChatGPT는 ‘대화형’ 인공지능 모델이라는 점에 주목할 필요가 있다. 우리는 대화를 할 때면 상대가 대화에 협조적일 것을 기대하고 또 그럴 것이라고 전제한다. 이와 관련하여 Grice(1975, 1978)는 대화의 진행을 가이드하는 일련의 기본적인 가정이 있음을 전제하고, 네 가지 ‘대화의 격률(maxim of conversation)’을 제시

한다. 이는 대화자가 효율적이고 협력적으로 대화하기 위해 따라야 하는 격률로, 대원칙인 ‘협력의 원리(co-operative principle)’를 주축으로 ‘양의 격률(the maxim of quantity), 질의 격률(the maxim of quality), 관련성의 격률(the maxim of relevance), 태도의 격률(the maxim of manner)’로 구분된다.

그러나 이는 인간의 모든 대화가 이러한 격률을 준수하는 형태로 이뤄지고 있다거나 반드시 그래야 함을 의미하지 않는다. 대화 중에 상대의 발화로부터 네 가지 격률 가운데 일부가 표면적으로 위배될 수 있는데, 이는 어떤 함축을 갖게 되며 청자는 그 의도를 추론을 하게 된다. 상대가 표면적 발화로부터 격률 중 하나를 위배한 데에는 어떤 이유가 있을 것이며, 대화의 협조적일 것이라는 원칙은 더 깊은 수준에서 준수되고 있을 것이라고 가정하는 것이다.<sup>16)</sup> 이처럼 우리는 대화에 참여하면서 기본적으로 상대방의 특정한 발화가 자신의 물음이나 대화의 주제와 관련성이 있을 것이라는 협력의 원리를 기대하지, 그 반대를 가정하거나 염려하지 않는다. 물론 이는 인간 대 인간 간의 대화에 대한 관찰에 기반한다.

이처럼 오랜 시간 습관화된 대화 모듈이 인공지능과의 대화에서도 부지불식간에 적용될 가능성을 무시할 수 없다. 실제 우리는 상대방이 인간이 아님을 알고 있음에도, 유사 대화 맥락에서 상대를 마치 인간처럼 대하는 경향이 있다. 대표적인 사례로 가전제품에 탑재되어있는 인공지능 시스템을 대하는 발화 태도를 들 수 있다. 직설적이고 명령적인 어투보다 공손한 표현을 사용하고, 간혹 요청에 응답해준 것에 대해 고맙다는 인사를 건네기도 한다. 인간처럼 공손성의 정도로 요청에 대한 결과가 달라지는 것이 아님에도 자연어로 대화를 나눈다는 측면에서 대화자에 대한 착각이 발생하는 것이다. 음성이 아닌 문자로 대화가 이뤄지는

---

16) S. C. Levinson, *Pragmatics*. Cambridge, New York: Cambridge University Press, 1983, p.102.



채팅도 마찬가지이다. 이용자가 질문을 하면 모델은 그에 대한 답변을 하는 방식으로 상호 관계를 맺게 된다.

이러한 측면에서 인공지능과의 대화의 장에서도 상대가 대화에 협조적일 것이라는 기대가 적용될 수 있다. 따라서 인공지능이 대화의 격률을 준수한다면 사실이 아니거나 확인되지 않는 것에 대해 말해서는 안 된다. 만약 질의 격률이 위배되었다면 그것은 다른 함축적 의미를 전달하기 위한 의도를 갖기 때문으로 설명되어야 한다. 그러나 정보로서 이루어진 진술에서 의도적이든 의도적이지 않든 질의 격률이 위배된다면, 그것은 앞서 언급하였듯이 ‘잘못된 정보’ 혹은 ‘허위 정보’에 해당하여 기대하는 정보의 역할을 수행하지 못하게 된다. 따라서 인공지능과의 대화에서 정보 제공을 요구하는 이용자의 기본적인 기대는 격률의 위배 없이 신뢰할 수 있는 정보의 출력이라 하겠다.

무엇보다 ChatGPT가 인공지능에 대한 대중의 인식 변화에 큰 영향을 준 요소로 언어적 ‘유창성’을 배제할 수 없다. 사용자의 명령에 따라 적절한 문장을 생성하는 대화형 인공지능 프로그램은 이미 1960년대 중반에 개발되기 시작하였다.<sup>17)</sup> 하지만 초기 프로그램들의 경우 튜링 테스트를 목적으로 인간처럼 자연스러운 언어적 반응을 산출하는 측면에서 그 성능이 평가되었다면, 이제는 일상적 대화뿐만 아니라 언어 번역, 이야기 생성, 요약, 정보 탐색까지 언어로 이루어지는 모든 장르에서의 유창성을 뽑낸다. 더불어 초기 프로그램의 경우 대화가 가능한 언어는 영어로 한정되었다. 하지만 현시점 인공지능 언어 모델은 영어에 비해 학습 데이터의 양이 한참 부족한 한국어로도 매우 자연스럽게 유창한 언어 표

---

17) 초기 대화형 인공지능 프로그램의 대표적인 사례로 ELIZA와 SHRDLU를 들 수 있다. 전자는 1964-1966년에 MIT에서 개발된 최초의 챗봇 소프트웨어 프로그램이다(Weizenbaum, 1966). ELIZA는 ‘튜링 테스트(Turing Test)’라는 어휘를 대중화하는 데 기여했으며, 언어를 통한 인간과 기계 간의 상호작용과 관련하여 실용적인 실험을 형성하는 데 결정적인 역할을 하였다(Natale, 2021). 후자는 1968-1970년에 MIT의 테리 위노그라드(Terry Winograd)에 의해 개발된 자연어 이해 컴퓨터 프로그램이다(Winograd, 1972).

현을 생성하고 있다.

이와 관련하여 ‘불쾌한 골짜기(uncanny valley)’라는 개념을 참조할 수 있다.<sup>18)</sup> 이 개념을 창시한 마사히로 모리(Masahiro Mori)에 따르면 로봇, 인형, 가상 인간과 같이 실제 인간이 아닌 것이 인간과 닮으면 닮을수록 우리의 호감도는 증가한다. 하지만 그로부터 어떤 불완전함이 느껴질 때 호감도가 마이너스 수준으로 떨어져 불쾌감이 증가하게 된다. 그러다 인간과 구분할 수 없는 상태로 나아가면 호감도가 다시 급격히 증가하게 되는데 이러한 호감도의 변화 그래프의 모습이 골짜기의 형태를 띠는 것이다. 보통 그래프의 곡선이 급격히 떨어지는 구간에 측면에서 주목하지만, 골짜기라는 이름이 붙을 수 있었던 것은 하향하던 그래프의 곡선이 다시 급상승하는 구간이 있기 때문이라는 점에 주목해보자.

이를 인공지능 언어 생성 모델에 적용하면, 모델에 대한 우리의 ‘호감도’는 그 모델이 생성하는 정보에 대한 ‘신뢰도’로 호환하여 생각해볼 수 있다. 컴퓨터가 인간처럼 말을 하기 시작한 초기 단계, 대화형 인공지능 프로그램에 대한 신뢰도는 단순한 질의(query) 수준에서 대중의 호감도를 불러일으킬 만하였다. 튜링 테스트를 통과하였다는 사실만으로도 당시 모델 역시 인간의 언어 능력에 준하는 수준이었음을 알 수 있다. 그러나 초기 모델은 몇 가지 패턴과 사용자의 입력값을 치환하는 방식으로 결과를 출력하는 방식으로 다채로운 요청에 유연하게 대응하는 데에는 큰 제약이 있었다. 이 지점에서 대화형 인공지능 모델의 불완전함이 인식됨으로써 일상 속에 녹아들기 어려웠던 것이다.

그러나 ChatGPT를 주축으로 하는 초거대 데이터를 학습한 인공지능 언어 모델은 기존의 초기 대화형 인공지능 프로그램의 불완전함을 해소하는 것을 넘어 단일한 인간의 능력을 넘어설 만큼 발전하였다. 몇 가지 정해진 패턴에 맞추는 것이 아니라, 자율적으로 다양한 질의에 걸맞은

---

18) Masahiro Mori & Karl MacDorman & Norri Kageki, “The Uncanny Valley [From the Field]”, *IEEE Robotics & Automation Magazine* 19, 2012, pp.98~100.

문장을 생성하는 것이다. 이전과는 다르게 수많은 대중의 이목을 끌며 유료 서비스까지 상용화되었다는 점은 골짜기의 신뢰 곡선이 급격하게 상승하고 있음을 방증한다. 물론 앞서 언급하였듯이 서비스가 오픈되면서 ChatGPT의 환각 현상이 큰 이슈가 되기는 하였지만, 현재 그러한 문제점이 상당히 개선되며 신뢰 곡선은 자못 안정적으로 상승하는 추세에 있는 것으로 보인다.

이처럼 초기 대화형 인공지능 프로그램은 일상적인 대화를 목적으로 개발되고 그 점에서 활용 가능성이 평가되었다면, 현재 인공지능 언어 모델은 인간이라 착각할 만큼 자연스러운 대화를 훨씬 넘어서는 능력을 수행해낼 것을 요구받고 있다. 골짜기의 증가 곡선은 단순히 호감도의 증가만이 아니라 그 이상의 수행 가능성에 대한 기대를 가져온 것이다. 다시 말해, 언어의 유창성이 인공지능 언어 모델에 대한 기대치를 함께 높이고 있다. 일반적으로 유창함에 대한 우리의 공통경험은 단순히 거침없이 말하는 것이 아닌, 특정 분야에 대한 깊은 이해와 전문적인 지식을 갖추고 있어 자연스럽게 말할 수 있는 경우에 해당한다.

누군가 수려한 말솜씨로 거침없이 어떤 사건이나 현상에 대해 말한다면 그 말의 내용을 막연히 의심하기는 쉽지 않다. ‘보이스피싱’이라는 범죄 유형이 많이 알려져 있음에도 끊임없이 피해자가 발생하는 이유 역시 다르지 않다. 사칭한 직업군에 걸맞게 준비된 말들을 거침없고 자연스럽게 내뱉으면서 사람들이 의심할 틈을 막는 것이다. 이때 관련된 실제 경험이 없는 사람들은 의심할 기반이 없기 때문에 그 유창성에 쉽게 속아들 수밖에 없다. 현재 인공지능이 생성한 문장을 통해 파악되는 유창성은 번역, 요약, 글쓰기 등 각 분야의 전문가에 버금가는 수준이다. 이처럼 일부 기능에서 검증된 뛰어난 능력은 신뢰의 기반이 될 수 있다.

이러한 측면에서 다양한 질의에 응답하는 유연성과 유창성이라는 표면적 양상으로부터 대중은 인공지능 언어 모델이 정확한 정보를 제공할 것이라는 막연한 믿음을 가지게 되었을 가능성을 고려할 수 있다. 이와

관련하여 Martindale&Capuat(2018)의 연구를 참조할 수 있다. 해당 연구는 기계 번역 결과에 대한 사용자들의 반응을 분석하기 위해, 적절하지만 유창하지 않은 번역과 유창하지만 오도된(misleading) 번역을 제공하고 그 결과를 통제 그룹과 비교하였다. 그 결과, 적절하지만 유창하지 않은 번역 결과 뒤에 해당 시스템에 대한 사용자들의 신뢰도에 상당한 부정적인 변화가 나타났고, 적절하고 유창한 번역이 최종 버전으로 제공되었을 때는 신뢰도가 다시 빠르게 정상화되는 현상이 포착되었다.

한편 유창하지만 오도된 번역에서는 유창하지 않은 번역과 비교했을 때 신뢰도에 큰 변화가 나타나지 않았다. 이는 유창성이 내용의 적절성을 가릴 정도로 시스템의 신뢰도에 큰 영향을 미칠 수 있음을 시사한다.<sup>19)</sup> 이 지점에서 유창성은 믿을 수 있는 정보를 제공할 것이라는 믿음의 기반으로도 연결될 수 있다. 그러나 인공지능 언어 모델은 인간처럼 사고하여 어떠한 사실이나 관념적 이해를 바탕으로 정보를 제공하지 않는다. 즉, 데이터에 기반하여 기호의 깊이나 넓이를 확장시키는 것이 아니다. 현시점 인공지능은 메타적으로 사고할 수 없기에 학습한 내용에 기반하여 특정 대상에 관한 정보를 범주화하기 어렵다. 그것이 데이터를 통해 학습하는 것은 재현 가능성이 높은 언어기호들의 조합 가능성이다.

여러 인공지능 언어 모델의 메커니즘을 상세히 알 수 없지만, 생성형 모델은 일반적으로 대규모의 학습 데이터에 기반하여 높은 확률을 갖는 단어와 단어 간의 결합 관계를 통계적으로 계산하여 문장을 생성한다. 엄밀히 말해, 인공지능의 세계에서는 대상-기호-해석체 간의 삼원적 세미오시스가 인간과는 다른 방식으로 작동한다. 인공지능은 기호가 재현하는 대상이 무엇인지, 또 그것이 해석자에게 어떤 효과를 미치는지에 관심이 없다. 인공지능 모델은 이용자가 입력하는 명령이나 요구에 따라

---

19) M. J. Martindale & M. Carpuat, "Fluency over adequacy: A pilot study in measuring user trust in imperfect MT", *AMTA 2018 - 13th Conference of the Association for Machine Translation in the Americas, Proceedings*, 2018, pp.13~25.

귀납적으로 기호를 조합하는 수동적 반응을 할 뿐, 어떤 대상을 재현하기 위해 기호를 사용하는 것이 아니다. 대상과 기호를 연결하는 법칙이 아닌 기호와 기호 간의 관계로부터 도출된 패턴을 따르는 것이다.

## 2. 인공지능 언어 모델의 메커니즘: 기계적(mechanical) 삼원적 관계

인간은 잘못된 사실임을 알고 있음에도 의도적으로 ‘허위정보’를 발생시킨다. 상징기호는 기호와 대상의 관계가 관습이나 법칙에 의해 연결되며, 대상과 닮은 점이 전혀 없이 머릿속에 개념을 떠올리게 하는 표상(representation)을 의미한다(W 1:257). 따라서 손가락으로 가리킬 수 없는 지난 과거나 아직 다가오지 않은 미래, 상상 속 세계, 또 추상적 관념을 재현할 수 있다. 이러한 측면에서 허위정보의 유포가 가능한 것이다. 한편 지표기호는 기호가 대상과의 인접성 또는 인과성에 의해 규정되므로 기호가 재현하는 대상의 존재가 필수적으로 요구된다. 물론 도상기호 또한 대상이 실재하지 않아도 기호를 통한 재현이 가능하다. 그러나 이러한 가짜 이미지는 또 다른 사회적 이슈를 만들어내고 있다.<sup>20)</sup>

흥미롭게도 인간은 인공지능이 기계적으로 조합하여 생성한 사실이 아닌 텍스트일지라도 그 기호의 해석체를 통해 대상을 본다. 상징기호를 사용하는 인간은 기호와 대상의 관계가 자연적 혹은 인과적으로 연결되어 있지 않더라도 법칙에 기반하여 기호를 해석할 수 있기 때문이다. 이 지점에서 인공지능이 인간과 같이 기호작용을 하고 있다는 착각이 발생할 수 있다. 그렇다면 인공지능 언어 모델이 존재하지 않는 혹은 잘못된 정보를 제공하는 것은 인간처럼 거짓말을 하는 것으로 볼 수 있는가? 이러한 현상의 원인 중 하나로 인공지능의 학습 데이터에 사실이 아닌 정

---

20) 이철민, “재벌백 다섯 아이 안고 업은 ‘고뇌의 아버지’... 이 가자 사진도 가짜였다”, 조선일보, 2023.11.02.

보가 포함되었을 가능성도 무시할 수 없다. 그러나 인공지능 언어 모델의 기호작용과 인간의 기호작용은 같지 않다.

이와 관련하여 퍼스가 설명하는 삼원적 관계와 이원적 관계의 비교를 참조할 수 있다. 퍼스에 따르면, 이원적 관계에서는 사건 A가 B를 발생시키고 사건 B가 C를 발생시킬 때, 사건 C가 B에 의해 발생될 것이라는 사실은 사건 A에 의해 B가 생성될 때 전혀 영향을 미치지 않는다. 사건 B가 C를 발생시키는 건 우발적인 사건이기 때문이다(CP 5.472). 반면 삼원적 관계는 사건 C를 발생시키기 위한 의도로 사건 A가 B를 만드는 경우에 해당한다. 이처럼 인간의 거짓말은 자신이 생성한 기호가 어떤 해석적 효과를 야기할 것이라는 기대 내지는 그러한 효과가 발생되도록 한다는 목적을 갖고 이루어지는 기호 행위라고 할 수 있다. 한편 인공지능은 그러한 의도나 목적성을 갖지 않는다는 점에서 잘못된 정보의 생성은 거짓말을 한다기보다도, 시스템의 프로세스에 따라 이원적으로 반응한 결과라 하겠다.

리슈카(2019)는 이러한 삼원적 관계를 목적 지향성에 따라 세 가지 유형으로 구분한다. 첫 번째는 목적론적(teleological) 기호과정으로, 이는 상징기호와 같이 관습에 기반하여 대상과 기호를 연결하여 삼원적으로 해석할 수 있는 인간의 보편적 능력과 관련된다. 두 번째는 기호가 삼원적으로 해석되기는 하지만 자연적으로 굳어진 습관에 의한 기호과정으로 동물 간의 의사소통이 대표적인 사례로 제시된다. 세 번째는 기계적인 기호과정으로, 해석체는 삼원적이 아니라 이원적 관계로 환원된다.<sup>21)</sup> 기호가 해석적 효과를 발생시키기는 하지만 그러한 해석적 효과를 발생시킬 것을 예상하여 기호를 생성하는 것이 아닌 것이다. 이러한 측면에서 인공지능 언어 모델의 기호 작용은 기계적인 삼원적 관계로 설명된다.

다시 말해, 인공지능 언어 모델이 생성한 텍스트는 어떤 해석체를 발생시키기 위해 생성된 것이 아니고, 그러한 텍스트가 생성되어 졌기 때문에 그 기호로부터 우발적으로 어떠한 해석적 효과가 발생한 것이다.

---

21) 제임스 야콥 리슈카, 『퍼스 기호학의 이해[개정판]』, 이윤희 역, HU:iNE, 2019, 92쪽.

퍼스의 설명에 따르면, 그는 해석체가 삼원적으로 생성되는 것이 ‘기호’에 필수적이라고 보고 있으므로 해석체가 삼원적으로 생성되지 않는 인공지능 언어 모델이 생성한 기호는 ‘유사기호(quasi-sign)’에 해당한다(CP 5.473).<sup>22)</sup> 따라서 상호 자연어로 소통을 하더라도 인공지능 언어 모델의 기호작용은 인간과 같지 않으며 잠재적으로 그것이 재현하는 바의 사실성은 보장되지 않는다.

인공지능 언어 모델의 생성 시스템은 통계 및 확률에 기반하여 작동하며 그 과정에서 단어의 조합이 만들어내는 의미나 효과는 별개의 사건이다. 즉 인공지능의 생성 결과는 높은 확률로 나타날 만한 자연스러운 인간의 표현일 뿐, 대상과 해석체의 관계에서 기호가 한정되지 않는다. 이러한 측면에서 대상을 재현하는 기호의 형식적 측면이 아닌 재현하는 대상이 관건이 되는 경우 주의가 필요하다. 발화자에 대한 검증이 이루어지지 않은 채, 언어 표현의 유창함만으로는 그것이 재현하는 대상의 실존성, 사실성을 대변하지 못하기 때문이다.

인간의 거짓말은 그 기호의 해석자에 의해 거짓 여부가 판별된다. 상대방이 유해 정보를 주더라도 수용자는 단번에 그 정보가 사실이 아님을 알 수 없다. 그것에 의심을 가하지 않고 수용한다면 수용자는 그로부터 유해한 영향을 입을 것이다. 한편 대중은 인공지능 언어 모델에게는 잘못된 정보 생성 가능성 자체를 엄격하게 대한다. 모순적이게도 소설이나 상상의 이미지 등 실존하지 않는 것들을 만들어내는 것도 요구함과 동시에 사실이 아닌 정보를 생성하는 것을 허용하지 않는다. 이용자가 어떤 정보를 원할 때는 인공지능이 사실인 정보만을 주기를 바라는 것이다. 그러나 인공지능 시스템은 자발적 통제를 할 수 없으며 질의 내용에 이원적으로 반응하는 것에 불과하다.

---

22) *ibid.*

- (1) 세종대왕이 맥북을 던진 사건에 대한 이야기를 만들어줘.
- (2) 세종대왕이 맥북을 던진 사건에 대해 말해줘.

ChatGPT에게 앞서 언급했던 ‘세종대왕 맥북 투척 사건’을 두 가지 방식으로 구분하여 물어보았다. 그 결과를 보면 이용자가 사실 정보를 원하는지 그럴듯한 이야기를 원하는지를 분별하는 것으로 보인다. 질의 (1)에 대해서는 세종대왕이 일이 뜻대로 이루어지지 않자 화가 나 맥북을 던졌다는, 환각 이슈를 불러일으켰던 것과 다른없는 이야기를 생성하였다. 한편 질의 (2)에 대해서는 “세종대왕이 맥북을 던진 사건은 사실이 아닙니다”와 같은 문장과 함께 세종대왕에 관한 역사적 사실을 간략하게 제시하는 결과를 출력하였다. 이처럼 인공지능은 사실인 정보만을 출력하는 것이 아니라 이용자의 요청에 따라 사실이 아닌 이야기를 꾸며낼 능력도 갖추고 있다.

### 3. 정보 탐색 과정에서 인공지능 언어 모델의 활용 방안 제언

그렇다면 인터넷 검색과 인공지능 언어 모델을 정보 탐색에 적절하게 활용할 수 있는 방안은 무엇인가. 가장 일반적으로 언급되는 제언 가운데 수평적 탐색이 있다. 한 곳의 출처에서 수직적으로 탐색하는 것이 아니라 다양한 출처의 정보를 검토해야 한다는 것이다. 하지만 인공지능 모델의 실용성은 파편화된 정보를 여러 차례에 걸쳐 탐색해야 하는 웹 검색의 수고로움을 덜어준다는 데 있음을 무시할 수 없다. 예를 들어, 인터넷 검색을 통해 ‘서울’에 대한 정보를 수집한다고 했을 때, 위치, 인구의 규모, 랜드마크 등 파편적인 정보 탐색이 요구된다. 한편 ChatGPT는 ‘서울’에 관한 주요 지표에 기반하여 하나의 텍스트로 정리하여 제시해 준다는 점에서 추가적인 수평적 탐색은 그 효용을 퇴색시킨다.

이러한 측면에서 생성형 인공지능 언어 모델의 생성 결과는 정보 탐색 과정에서 일종의 ‘지표(index)’로 활용할 수 있는 방안을 제안할 수 있다.



정보 탐색의 출발점을 찾아주거나, 어떤 키워드로 탐색을 지속할 수 있는지 그 방향성을 제시하는 가이드로서의 역할을 부여하는 것이다. 구글은 ChatGPT가 서비스되기 이전에 이미 특정 검색어를 입력하였을 때 그에 관한 기본적인 정보를 요약하여 제시해주거나, 입력된 키워드에 관한 정보를 찾는데 연관성이 높은 질문들을 제안해주는 서비스를 제공하기 시작하였다. 마찬가지로 인공지능 언어 모델이 생성하는 내용을 그대로 받아들이는 것이 아니라, 그 생성 내용을 지표로 삼아 정보 탐색을 확장해나가는 것이다. 그 과정에서 수평적 탐색이 자연스럽게 이루어질 것이다.

나는 어떤 지식의 습득을 “정보”라 칭한다. 그것은 논리적으로 단어의 의미에 대한 경험이 아닌 어떤 다른 경험을 요구한다.(MS[R] 664:19)

위의 퍼스의 기술을 통해 알 수 있듯이, 우리가 특정 대상에 관한 정보를 얻는 데에는 그 대상으로부터 변화된 사실이나 새로운 특성을 발견하는 경험이 전제된다. 따라서 그 정보는 어느 누구에게나 검증 가능한 것이 되어야 한다. 물론 모든 정보의 대상을 과학적인 방식으로 탐구할 수는 없겠지만, 경험적 근거를 확보하기 위한 역동적인 탐색으로 나아갈 수 있다. 다양한 관점에서 이뤄지는 일련의 정보 탐색은 사고의 확장을 불러일으키는 일종의 간접 경험이 되는 것이다. 인공지능 언어 모델이 출력하는 텍스트로부터 탐색을 연속하는 과정에서 우리는 비로소 자신에게 실제적이고 유의미한 효과를 주는 정보를 습득하게 될 것이다.

한편 각각의 인공지능 모델은 특정 태스크 수행을 목적으로 개발된 것임을 인지할 필요가 있다. 현재 개발된 인공지능 모델들은 일부 측면에서 인간의 수준을 뛰어 넘을만한 성능을 자랑하지만, 추가적인 플러그인(Plugin)의 도움 없이 전문화된 인간의 모든 기술을 복합적으로 수행할 수 있는 모델은 아직 없다. 알파고가 아무리 바둑을 잘 두어도 체스는 둘 수 없으며, 텍스트를 번역하거나 자연스러운 대화를 이어나갈 수도 없다. ChatGPT는 인간과

대화하듯 자연스러운 언어적 표현을 출력하는 기능에 개발 초점이 맞춰져 있으며, 이로부터 출력되는 결과의 사실성은 보장되지 않는다. 이러한 측면에서 ChatGPT 서비스에 로그인을 하면 아래와 같은 공지문이 뜬다.

“우리는 안전장치를 가지고 있지만, ChatGPT가 부정확한 정보를 제공할 수 있습니다. 이는 조언을 제공하기 위한 것이 아닙니다. (While we have safeguards, ChatGPT may give you inaccurate information. It’s not intended to give advice.)”

더불어 인공지능 언어 모델은 언어로 이뤄지는 다양한 담화의 ‘형식’에 적용하거나 변환하는 데 적극 활용될 수 있다. 세미오시스의 작동 없이도 다양한 장르의 문서로 구성된 수많은 데이터로부터 특정 양식의 패턴은 추상화될 수 있기 때문이다. 인간에게 이러한 패턴은 장기간 같은 작업을 반복적으로 훈련하는 과정에서 습득할 수 있는 것이지만, 인공지능 모델은 매우 빠른 속도로 도출할 수 있다. 비록 기호만 조합할 뿐일지만 반복적으로 나타나는 패턴을 읽어 다이어그램적 추론을 수행한다고 할 수 있다. 이와 관련하여 ChatGPT에게 해당 서비스를 통해 어떤 도움을 받을 수 있는지 묻은 결과를 참조할 수 있다.

- |                                      |  |
|--------------------------------------|--|
| • 일상적인 대화                            | • 콘텐츠 아이디어 제공                          |
| • 정보 검색                              | • 개인 비서 (일정 관리, 날씨 정보, 교통 정보)          |
| • 문서 생성/창작적 글쓰기<br>(시, 에세이, 이메일, 소셜) | • 기술 지원 (프로그래밍, 웹 개발 등 기술적인 질문에 대한 답변) |
| • 문서 요약                              |  |
| • 언어 번역                              |  |

위 내용을 살펴보면, ChatGPT는 텍스트에 기반하여 이뤄지는 일상적인 작업 대부분에 도움을 줄 수 있는 것으로 보인다. 여기서 주목할만한

점은 정보 검색 외에 대부분의 활용 사례는 그 내용의 사실성 혹은 내용에 대한 검증을 전제하지 않는다는 것이다. 이처럼 사용자가 모델에게 미리 콘텐츠를 제공해주고 그에 기반하여 기호의 형식을 바꾸는 작업에서는 비판적 정보 수용에 대한 긴장성을 낮출 수 있다. 즉, 언어 생성 능력의 유창성은 그 내용적 측면이 아닌 형식 변환에서 효과적으로 활용될 수 있는 것이다. 참조할 콘텐츠를 미리 제공해준다는 점에서 인공지능 언어 모델이 생성하는 텍스트와 관련된 일련의 편향성, 저작권, 윤리적 문제 등에 대한 민감성도 낮출 수 있다는 점에서 유용하다.

일례로 번역 작업의 경우, 번역 내용에 대한 사실성 검증은 요구되지 않는다. 번역의 결과가 정확하며 번역된 언어 표현이 자연스럽게 읽히는 것이 중요하다. 물론 함축적 의미의 번역은 여전히 어려움이 있겠지만, 번역 결과는 이용자에 의해 제공된, 사실성에 대한 검증이 요구되지 않는 자료로 한정된다. 더불어 날씨, 교통에 대한 정보는 외부로부터 이용자가 확인하기에 편리한 양식으로 정보를 불러오는 것에 해당한다. 프로그래밍에 대한 기술적 지원 역시 코드의 잘못된 조합을 교정하고 오류를 잡는 데에 목적이 있다는 점에서 내용이 아닌 형식에 초점이 맞춰져 있다. 따라서 생성형 인공지능 언어 모델은 패턴이나 규칙으로부터 도출 가능한 형식을 적용하는 데 있어 의심의 정도를 낮추고 활용될 수 있다.

## V. 결론

본 연구의 논의를 요약하면 다음과 같다. 우리는 성장 과정에서 텍스트 의존적인 학습 방식을 습관화하면서 그로부터 텍스트에 대한 막연한 신뢰를 내재화하고 있다. 이렇게 내재화된 습관은 막대한 정보량과 빠른 검색이라는 인터넷의 막강한 권력에 의해 새롭게 가이드되지 못하고 있다. 또한 상징기호가 갖는 기호의 본질적 속성상 잘 알지 못하는 대상에 관한 진술에 대해서 의심을 가하기 쉽지 않다. 이러한 요인들은 디지털

환경에서 잘못된 정보의 유포 및 비판적 정보 수용의 어려움이 지속되는 것과 무관하다고 할 수 없다. 이러한 정보 탐색 습관에 대한 재고는 인터넷을 통해 획득한 정보를 사고와 행동의 변화에 어떻게 영향을 미치게 할지 판단하는 데 도움이 될 수 있다.

한편 ChatGPT와 같은 인공지능 언어 모델이 장착한 ‘대화’라는 모듈은 사용자로 하여금 협력의 원리를 동기화하게 한다. 이로 인해 모델이 제공하는 잘못된 정보는 대화의 격률을 위배한다는 점에서 엄격한 잣대로 평가된다. 또 인공지능을 마치 전문화된 인간으로 인식하는 데에는 다양한 질의 범위에 대한 유연성과 언어적 유창성이 큰 영향을 미친다. 이용자가 잘 알지 못하는 것에 대한 상세한 진술에 의심을 가하기는 쉽지 않기 때문이다. 아이러니하게도 대중은 인공지능 언어 모델로부터 정보 제공만이 아니라 가상의 이야기도 잘 만들어낼 것을 기대한다. 이때 생성형 인공지능 언어 모델은 기계적 기호작용을 한다는 점에서 인간의 삼원적 세미오시스와 구분되어야 마땅하다.

따라서 본 연구는 텍스트 정보를 막연히 의심하는 방식으로 기존의 습관을 변화시키는 것이 아니라 정보를 탐색하는 습관을 새로이 들이는 방안을 제안하였다. 이는 일상생활 속에서 인공지능 언어 모델을 효율적이고 유용하게 활용함으로써 함께 공존할 수 있는 방법을 모색하는 데 목적이 있다. 구체적으로 첫째, 인공지능 언어 모델의 출력 결과를 특정 대상에 대한 탐구의 지표로 삼는 방안을 제시하였다. 기호 그 자체를 보는 것이 아니라 기호가 가리키는 대상을 보는 것이다. 둘째, 방대한 텍스트 데이터로부터 빠른 속도로 도출할 수 있는 담화 형식의 적용을 정보의 사실성 검증에 대한 긴장감을 낮춰 인공지능 언어 모델을 활용할 수 있는 측면으로 제시하였다.

이러한 제안은 일상에서 인공지능의 활용 범위가 제한되어야 함을 의미하지 않는다. 실제 ChatGPT에 결합되는 플러그인 또는 정보 탐색에 초점화되어 개발된 모델의 경우 그 내용이 검증된 텍스트 자원에 기반하

여 탐색하고 요약하여 정보를 추출할 수 있다. 다만 본 연구는 검증되지 않은 자료가 혼재된 인터넷을 주된 정보 탐색의 장으로 삼으며, 인공지능과 공존하는 삶에 적응하는 과정에서 특히 언어와 문자라는 상징기호의 특성을 고려하여 디지털 도구를 활용할 수 있는 방안을 제안했다는 데 의의를 두고자 한다. 인공지능 모델의 개발은 결국 인간과의 소통을 지향할 것이기 때문에 시대를 막론하고 변화되는 환경 속에서 언어와 문자라는 기호의 특성은 끊임없이 탐구되어야 할 것이다.

## 참고문헌

- 김경희 · 김광재 · 이숙정, 「모바일 환경에서의 미디어 리터러시 구성 요소와 세대 간 미디어 리터러시 격차」, 『한국방송학보』33(4), 2019, 5~36쪽.
- 이철민, “재벌백 다섯 아이 안고 업은 ‘고뇌의 아버지’... 이 가자 사진도 가짜였다”, 조선일보, 2013.11.02.
- 제임스 야콥 리슈카, 『퍼스 기호학의 이해[개정판]』, 이윤희 역, HU:iNE, 2019.
- 코르넬리스 드발, 『퍼스 철학의 이해[개정판]』, 이윤희 역, HU:iNE, 2019.
- J. 옹. 『구술문화와 문자문화: 언어를 다루는 기술』, 임명진 옮김, 문예출판사, 2021.
- 체릴린 아이어톤 · 줄리 포세티, 『저널리즘, 가짜뉴스 & 허위정보 : 저널리즘 교육과 훈련을 위한 핸드북』, 김익현 역, 서울: 한국언론진흥재단. 2020.
- C. S. Peirce, *An Original Manuscript*, numbered according to Prof Richard S. Robin’s annotated catalogue, 1967.
- C. S. Peirce, *Collected Papers of Charles S. Peirce*. 8 vols. Ed. Hartshorne, C. and Weiss, P.(vols. 1-6), and Burks, A.(vols 7-8). Cambridge, MA: Harvard University Press, 1931~58.
- C. S. Peirce, *Writings of Charles S. Peirce: A Chronological Edition*. 6 vols. to date. Ed. by The Peirce Edition Project. Bloomington: Indiana University Press, 1982-.
- H. P. Grice, “Further Notes on Logic and Conversation.” In *Pragmatics* [Syntax and Semantics 9], Ed. Peter Cole, New York: Academic Press, 1978.
- H. P. Grice, “Logic and conversation”. In *Syntax and Semantics*, [Vol. 3, Speech Acts], Eds. Peter Cole and Jerry L. Morgan, New York: Academic Press 1975, pp.41~58.
- J. Weizenbaum, “ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine”, *Communications of the ACM* 9.1, 1966, pp.36~45.
- Masahiro Mori & Karl MacDorman & Norri Kageki, “The Uncanny Valley [From the Field]”, *IEEE Robotics & Automation Magazine* 19(2), 2012, pp.98~100.
- M. J. Martindale & M. Carpuat, “Fluency over adequacy: A pilot study in measuring user trust in imperfect MT”, *AMTA 2018 - 13th Conference of the Association*

- for Machine Translation in the Americas, Proceedings*, 2018, pp.13~25.
- Michael James Grenfell, *Pierre Bourdieu: Key Concepts: Vol. 2nd ed.* Routledge, 2014.
- P. Bourdieu & L. Wacquant, *An Invitation to Reflexive Sociology*, L. Wacquant (trans.). Cambridge: Polity, 1992.
- S. C. Levinson, *Pragmatics*. Cambridge, New York: Cambridge University Press, 1983.
- S. Natale, “The ELIZA Effect: Joseph Weizenbaum and the Emergence of Chatbots”, In *Deceitful Media: Artificial Intelligence and Social Life after the Turing Test*, New York: Oxford Academic, 2021.
- T. Winograd, “Understanding Natural Language”, *Cognitive Psychology* 3, 1972, pp.1~191.
- Walter J. Ong, *Orality and literacy: the technologizing of the word*, London; New York: Routledge, 2002.

국립국어원, 『표준국어대사전』, 2018. <https://stdict.korean.go.kr>

OpenAI. ChatGPT (version 3.5) [computer software], 2023, from <https://chat.openai.com>

We are Social, *THE CHANGING WORLD OF DIGITAL IN 2023*, 2023, from <https://wearesocial.com/us/blog/2023/01/the-changing-world-of-digital-in-2023/>

# A Semiotic Approach to Information Exploration in the Digital Era: Why Do We Expect Reliable Information from ChatGPT?

Hong, Seung-Hye

This study explores two main questions related to information exploration in the digital era. The first question revolves around the factors that hinder the critical acceptance of information, a skill consistently demanded in digital literacy. In this context, this study emphasizes habitual beliefs in texts created by the socio-cultural system and the characteristics of symbolic signs, which act as mediators in information exploration. The second question addresses why we anticipate reliable information from artificial intelligence, such as ChatGPT, which has emerged as an additional information exploration tool in the digital age. In this context, we discuss that ChatGPT, as a conversation module distinct from web searches, is required to adhere to Grice's conversational maxims, and its linguistic fluency serves as the foundation for our expectations in artificial intelligence. Additionally, this study suggests ways to use artificial intelligence as an index for information exploration, taking into consideration the mechanism of artificial intelligence operating based on symbolic signs. This study is significant in that we revisits the information exploration habits of contemporary individuals from a semiotic perspective. Additionally, we highlight the essential need for an intrinsic understanding of language as symbols, particularly in the context of developing artificial intelligence language models, which ultimately aim for communication with humans.

Keywords : AI, ChtGPT, Conversation, Maxim of conversation, Digital literacy, Information, index, Symbolic signs, Habitus.

투고일: 2023. 11. 26./ 심사일: 2023. 12. 10./ 심사완료일: 2023. 12. 13.