AI·Digital 이슈
공론화 리포트 2호

NIA
한국지능정보사회진흥원

AI 환각 : 위험 관리와 창의적 활용을 위한 정책 대응 전략

- AI 환각 관련 동향 및 정책적 시사점을 중심으로 -

디지털포용본부 디지털포용기획팀
고현지 선임(ghjnia@nia.or.kr)

「AI·Digital 이슈 공론화 Report」는 인공지능(AI)과 디지털 기술의 발전이 야기한 사회적 변화와 이슈를 다각도로 분석하고, AI 이슈에 대한 국민 인식 확산과 다층적인 정책 방향 제시를 목적으로 하는 한국지능정보사회진흥원의 보고서입니다.

■ 발 행 일 : 2025.12.3
■ 발 행 처 : 한국지능정보사회진흥원
■ 발 행 인 : 황 종 성
■ 기획·자문 : 디지털포용본부 최문실 본부장, 정기호 팀장
■ 작성 : 디지털포용본부 디지털포용기획팀 고현지 선임
■ 검수 : KAIST 김정남 교수, 성균관대 박종현 교수, 세종대 장윤 교수, KAIST DaLI Lab 서성범 연구원, 한국전자통신연구원 인공지능 안전연구소 송경호 선임연구원
■ 보고서 온라인 서비스 : www.nia.or.kr

Contents

NATIONAL INFORMATION SOCIETY AGENCY

AI 환각 : 위험 관리와 창의적 활용을 위한 정책 대응 전략
– AI 환각 관련 동향 및 정책적 시사점을 중심으로 –

PART I

서론

1 AI 시대의 위험인가 기회인가? : AI 환각

- 인공지능(AI) 기술, 특히 대규모 언어모델(LLM)과 생성형 AI의 발전은 정보생산 및 처리 방식에 혁신적 변화를 가져옴으로써, 사회전반의 효율성과 창의성을 증진시키는 긍정적 효과를 창출
- AI 활용의 대중화 시대가 열리고 AI 환각 현상은 단순한 오류를 넘어 AI의 영향권에 있는 모든 사회 시스템에 부정적 영향을 미칠 수 있는 위험으로 인식되며, AI 논의의 핵심 쟁점으로 부상
- 특히 AI에 대한 과도한 신뢰, 소위 'AI 만능주의'는 사용자와 사회가 AI에 생성된 정보를 무조건 정확하다고 받아들이게 만들며, AI 환각현상과 결합되는 경우 사회 전반의 신뢰 붕괴를 야기
- 실제로 호주 빅토리아주 대법원에서 AI가 만들어낸 허위 판례와 가짜 인용문이 제출된 사건[1]과 함께('25), 미국 플로리다 중부 지방 법원에서는 허위 판례를 인용한 변호사에게 정직 처분('23)[2]
- 미시건대학교 연구에 따르면 오픈 AI의 음성-텍스트 변환 AI모델 '위스퍼(Whisper)'의 경우, 10건의 오디오 필사본 중 8건에서 환각, 즉 존재하지 않는 내용을 생성하는 현상을 발견('24)[3]
- 이처럼 환각은 AI의 신뢰성 측면에서 단순한 오류가 아닌 법률, 의료 등 다양한 분야에서의 개인의 권익 침해, 공공 신뢰의 훼손 등으로 사회 전반에 부정적인 영향을 미칠 가능성
- 한편, 예술, 디자인, 문학 창작 분야에서는 AI 환각 현상이 사실을 넘어서는 창의적 발상과 혁신적 창출의 원천으로 활용될 수 있으며, 창의성을 촉진하는 도구로 평가받는 중
- '24년 노벨화학상 수상자인 데이비드 베이커(David Baker) 교수는 AI 환각이 자연에 존재하지 않는 1000만개의 새로운 단백질을 만드는데 기여하고, 핵심적인 역할을 했다고 언급[4]
- 이와 같이 AI 환각이 사회적 신뢰의 위협과 창의의 원천이라는 양면성을 지닌 만큼 사회적 신뢰 보호와 창의성 증진을 동시에 달성할 수 있는 정책적 대응이 요구되는 상황
- 따라서 본 보고서는 환각의 개념 및 유형, 사례 분석 및 정책 동향 파악을 통해, AI 환각에 대한 위험 최소화와 창의적 잠재력 제고를 모두 고려한 정책적 시사점을 도출하고자 함

PART Ⅱ

AI 환각의 개념과 유형

1 AI 환각의 개념

○ AI 환각(AI Hallucination)이란

- AI 환각(AI Hallucination)이란 인간의 인지적 착각(Hallucination)에서 유래된 용어로, 인공지능이 실제로 존재하지 않거나, 사실과 다른 정보를 진짜처럼 생성하는 오류 현상을 의미
- 그러나 인간의 인지적 착각과 달리, AI의 경우 확률 기반의 언어·이미지 생성 모델이 불완전한 학습 데이터나 문맥 예측의 불확실성에 의해 발생한다는 점에서 인간과의 구조적 차이가 존재

〈 기업별 AI 환각에 대한 정의 〉

기 업	정 의
OpenAI	- 언어모델이 생성하는 '그럴듯하지만 잘못된 진술'(plausible but false statements)[5]
Google	- AI 모델이 생성하는 부정확하거나 오해의 소지가 있는 결과(incorrect or misleading results)[6]
Microsoft	- AI에 의해 생성된 근거없는 콘텐츠("ungrounded" content)를 의미하며, 즉 모델이 주어진 데이터 자체를 변형하거나 데이터에 포함되지 않은 추가 정보를 만든 경우를 지칭[7]
IBM	- 대형 언어모델(LLM)이 존재하지 않는 패턴이나 객체를 인식·출력하는 현상(when a large language model (LLM) perceives patterns or objects that are nonexistent)[8]

- 초기 생성형 AI 모델의 경우 이러한 환각 현상이 심각했으며, 이러한 현상으로 인해 다양한 밈(Meme)*까지 확산, 최근 모델은 보다 개선된 경향을 보이며 정확도 및 신뢰성이 향상된 모습

* 세종대왕 맥북 던짐 사건 : AI가 '세종대왕이 맥북프로를 던졌다'는 허위 정보를 생성한 사례로 해당 밈은 인공지능의 환각 현상을 대중적으로 인식시키는 계기가 되었으며, 당시 AI 모델의 한계를 드러냄과 동시에 AI 신뢰성 확보에 대한 논의를 촉발

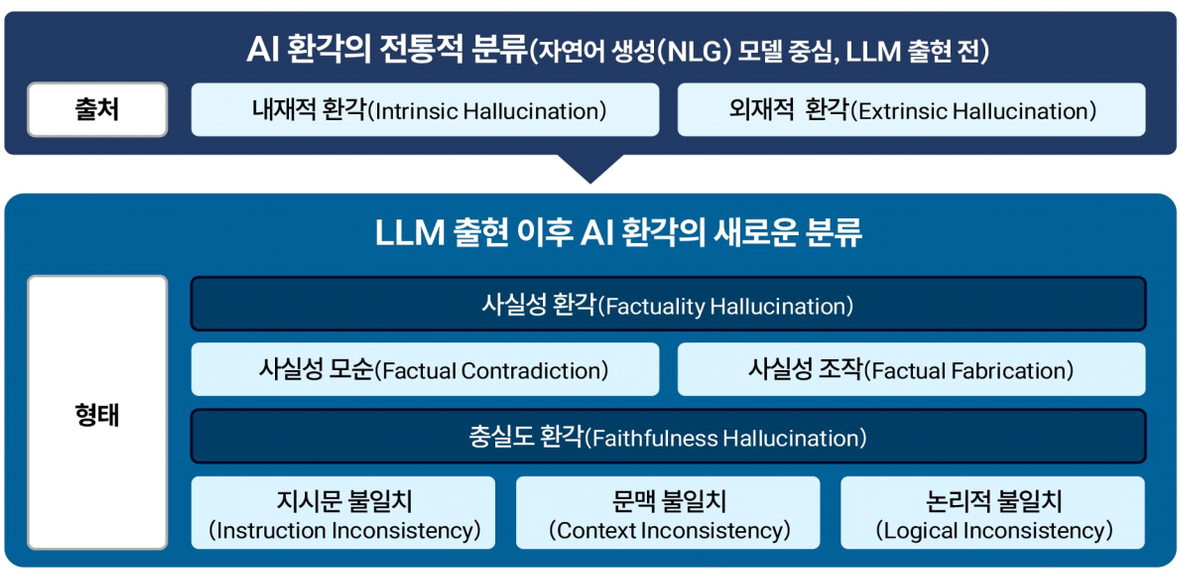
'세종대왕 맥북 던짐 사건'에 대한 GPT 입력 결과물('23)[9]	'세종대왕 맥북 던짐 사건'에 대한 GPT 입력 결과물('25)
조선왕조실록에 기록 된 세종대왕 맥북프로 던짐 사건에 대해 알려줘	세종대왕 맥북프로 던짐 사건에 대해서 알려줘
조선왕조실록에 기록된 세종대왕 맥북프로 던짐 사건은 2009년 1월 19일에 일어난 사건으로, 인터넷과 관련된 문제로 인해 불만을 가진 세종대왕이 맥북프로를 던졌다는 이야기입니다.	"세종대왕 맥북프로 던짐 사건" 이야기는 실제 역사적 사실이 아닙니다. 이야기의 기원은 인공지능 챗봇이 만들어 낸 허위 정보 또는 "환각(hallucination)" 답변 중 하나입니다. 국제신문 +2

2 AI 환각의 분류 및 유형

◎ 본 장에서는 환각 분야에서 다수 인용되는 'A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions'와 서비스 관점에서 환각을 분류한 NTT의 'All Hallucinations are Not Bad. Acknowledging Gen AI's Constraints and Benefits' 기준으로 환각 유형 및 발생원인을 소개[10]

- 현재 AI 환각에 대한 분류·유형은 공식적인 국제 표준이나 단일 분류 체계가 부재한 상황, 다양한 국가의 연구팀과 기관에서 제시한 분류체계가 있으나, 문헌마다 일부 내용의 차이가 존재

〈 AI 환각의 분류 및 유형 〉



○ 출처 기반 분류(전통적 분류)

- AI 모델에게 답변의 근거가 될 만한 소스 콘텐츠가 주어졌는가에 따라 환각은 내재적 환각(Intrinsic Hallucination)과 외재적 환각(Extrinsic Hallucination)으로 분류 가능[11]
- **내재적 환각 :** AI모델 답변의 근거가 될 만한 자료를 제공했을 때, 자료를 잘못 해석하거나 계산 오류를 일으키는 경우, 모델이 제시된 컨텍스트를 무시하고 모순되는 정보를 생성할 때 발생
  - **예시)** 사용자에게 나일강이 '중앙아프리카의 대호수 지역'에서 발원한다는 컨텍스트가 제공되었음에도 불구하고, 모델이 요약문에서 나일강이 '산맥'에서 발원한다고 응답
- **외재적 환각 :** AI 모델 답변의 근거가 될 만한 자료가 제공되지 않았거나 자료 밖의 지식을 사용할 때, AI 모델의 답변 내용이 외부 세계의 객관적 사실과 모순되는 경우 발생
  - **예시)** AI 모델이 "에펠탑 건설이 환경에 미친 영향"에 대해 '파리 호랑이(Parisian tiger)'의 멸종을 초래했다고 응답(자료가 제공되지 않아 근거를 찾을 수 없고 실제 사실과 모순되는 응답)

- 그러나 입력된 자료의 내용과 다르면서 현실의 사실과도 다른 경우에는 내재적, 외재적 둘 다 해당되므로, 두 분류가 명확하지 않고 중첩되며 작업 유형에 따라 구분이 모호하다는 지적

○ 형태 기반 분류(LLM 등장 이후 분류)

- 또한 자연어 생성(NLG) 모델에서 LLM이 등장하며 언어의 이해, 생성, 추론 등의 기능이 확대되면서 기존 환각 분류(내재적/외재적)의 한계를 인식한 연구진들은 분류 체계를 보다 확장[12)]
- 이에 LLM의 환각은 일반적으로 사실성 환각과 충실성 환각이라는 두 가지 주요 유형으로 분류, 이는 전통적 분류(내재적/외재적)의 개념과 일부 중첩되는 분류(상호배타적인 개념이 아님)
- **사실성 환각(Factuality Hallucination) :** 자료 제공과 관계없이 LLM이 생성한 내용이 실제 세계의 사실과 불일치하는 경우로, '사실적 모순'과 '사실 조작' 2개 유형으로 구분
- 사실적 모순은 LLM의 출력내용이 실제로 존재하는 확립된 사실을 명백히 부정하거나 사실과 모순되는 경우를 의미하며 개체 오류 환각, 관계 오류 환각 2개의 하위 세부 유형으로 구분

〈 사실적 모순의 세부 유형 〉

분 류	유 형	설명 및 예시
사실적 모순 (Factual Contradiction)	개체 오류 환각 (Entity-error hallucination)	- 생성된 텍스트에 잘못된 개체(erroneous entities)가 포함된 상황을 의미 - 예) 전화기 발명가를 '토마스 에디슨'이라고 잘못 답변
	관계 오류 환각 (Relation-error hallucination)	- 생성된 텍스트에 개체 간의 잘못된 관계(wrong relations)가 포함된 경우 - 예) 에디슨이 전구를 '발명'했다고 답변(실제로는 기존 디자인을 '개선'함)

- 사실 조작은 LLM의 출력 내용에 실제 존재하는 확립된 사실에 비추어 검증할 수 없는 사실이 포함되어 있을 때 발생하며, 출력 정보가 허구적이거나 근거가 없는 것을 의미

〈 사실 조작의 세부 유형 〉

분 류	유 형	설명 및 예시
사실 조작 (Factual Fabrication)	검증 불가능성 환각 (Unverifiability hallucination)	- 전적으로 존재하지 않거나 사용 가능한 출처를 통해 검증될 수 없는 진술을 의미 - 예) "에펠탑 건설이 파리 호랑이(Parisian tiger)의 멸종을 초래했다"는 AI 답변은 실제로 존재하지 않는 사실이며, 입증될 수 없는 날조된 주장
	과장된 주장 환각 (Overclaim hallucination)	- 주관적인 편향으로 인해 보편적인 타당성이 부족한 주장을 포함하는 경우 - 예) "에펠탑 건설이 전 세계 녹색 건축 운동을 촉발한 사건으로 널리 인정받고 있다"고 답변, 이는 광범위한 합의나 실질적인 증거가 부족한 과장된 주장

- **충실성 환각(Faithfulness Hallucination) :** 생성된 내용이 사용자 입력으로부터 벗어나거나, 생성된 내용 내부에 자체적인 일관성이 부족한 경우를 의미하며 총 3가지 세부 유형으로 분류

〈 충실성 환각의 세부 유형 〉

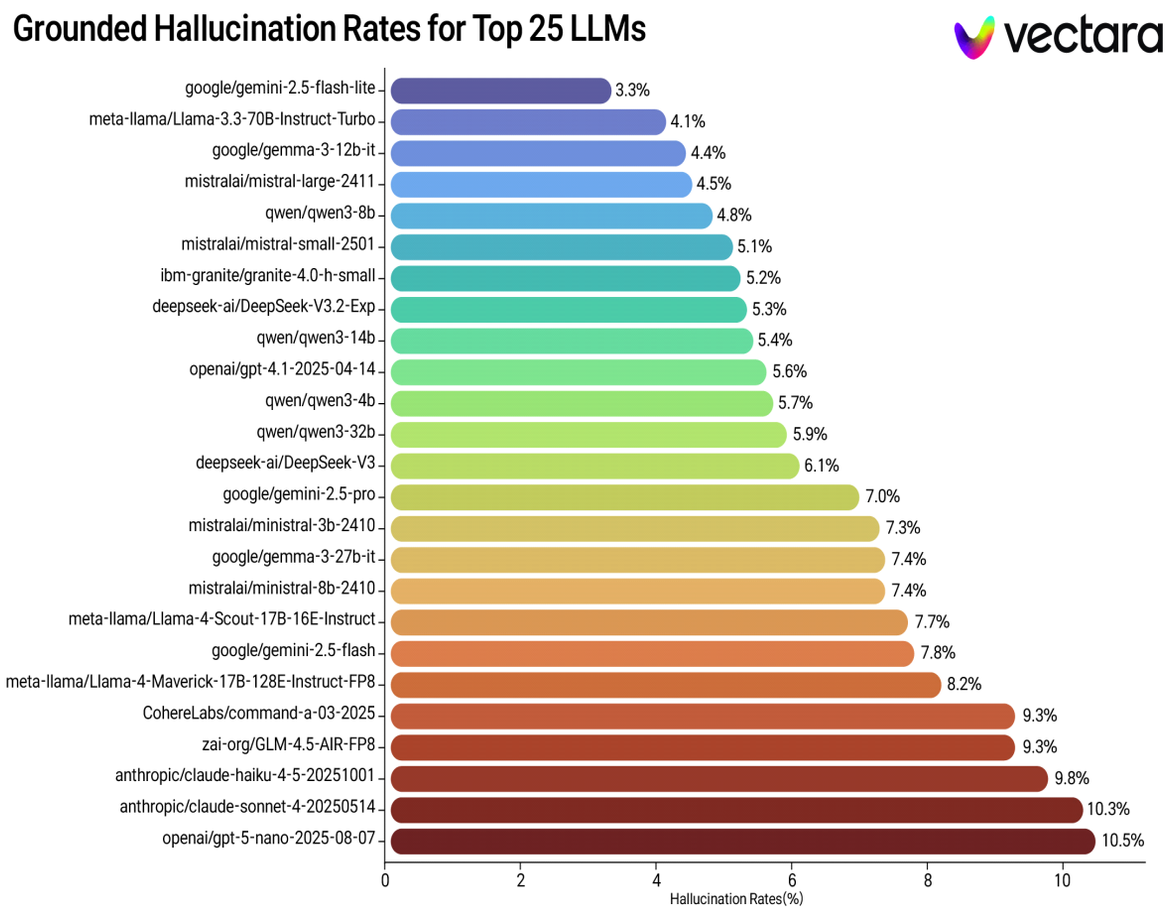
분 류	설명 및 예시
지시 불일치 (Instruction inconsistency)	- LLM의 출력이 사용자의 지시 사항에서 벗어나는 경우. 의도하지 않은 불일치를 의미 - 예) 사용자가 영어 질문을 스페인어로 번역하라고 지시했으나, 모델이 질문에 대한 답변을 하는 경우
맥락 불일치 (Context inconsistency)	- LLM의 출력이 사용자가 제공한 맥락 정보에 불충실한 경우 - 예) 사용자가 나일강의 발원지가 '중앙 아프리카의 오대호 지역'이라고 제공했으나, 모델은 '중앙 아프리카의 산맥'이라고 답변하는 경우
논리적 불일치 (Logical inconsistency)	- LLM의 출력이 내부적으로 논리적 모순을 보이는 경우. 특히 추론 작업에서 자주 관찰되며, 추론 단계 자체와 최종 답변 간의 불일치가 확인되는 경우 - 예) 방정식 풀이 단계는 맞았으나, 최종 답변이 앞선 추론 단계와 일관되지 않게 잘못 제시되는 경우

○ 사용 사례 기반 분류

- 일본의 최대 통신 기업 NTT 그룹의 자회사인 NTT 데이터에서는 AI 환각의 유형을 모델의 아키텍처 또는 의도된 사용 사례에 따라 분류, 이는 학문적 기준이 아닌 산업계의 기준으로서 의미를 지님[13)]
- **시각적 환각(Visual Hallucination) :** 이미지 생성 모델에서 존재하지 않는 개체, 장면, 패턴을 묘사하는 이미지를 생성, 초현실적이고 추상적인 예술작품부터 완전히 조작된 물체까지 다양
- **텍스트 환각(Textual Hallucination) :** 언어 모델이 허구적인 정보를 포함하거나 거짓 주장을 하는 문장·단락을 생성하는 것으로 근거가 없는 사건, 세부정보 혹은 사실을 날조하는 것을 포함
- **콘텐츠 확장 환각(Content Expansion Hallucination) :** 생성모델이 입력 데이터에 존재하는 정보보다 더 많은 정보를 생성할 때 실제로 존재하지 않는 내용을 확장해서 만들어 내는 현상
- **추론 환각(Inference Hallucination) :** 대규모 언어 모델(LLM)이 입력 데이터로부터 부정확한 가정이나 추론을 초래하여, 맥락을 오판하거나 잘못 표현하는 현상
- **편향 환각(Bias Hallucination) :** 학습데이터에 이미 존재하는 편향을 반영하거나 증폭시키는 콘텐츠를 생성하는 것을 의미, 이는 고정관념, 차별 등 비윤리적 관점을 나타내는 결과물을 초래
- **맥락적 환각(Contextual Hallucination) :** 언어모델이 맥락적으로는 관련성이 있는 것처럼 보이지만 실제로는 부정확하거나 실제 맥락을 대표하지 않는 텍스트를 생성

AI 모델별 환각률

- 최근 몇 년간 대규모 언어 모델(LLM)은 정보검색, 콘텐츠 생성, 자동화된 고객 응대 등 광범위한 분야에서 혁신을 주도하였으나, LLM의 환각 문제는 AI 도입의 큰 장애물로 인식
- 신뢰 기반 생성형 AI 플랫폼 기업인 Vectara는 LLM 환각률 리더보드를 통해 주요 모델들의 환각률을 비교하고, 신뢰도 높은 AI 시스템을 구축하기 위한 조사를 추진('25.11 기준)



- **(환각률 3%~4%)** Google Gemini-2.5 Flash Lite는 3.3%라는 극히 낮은 환각률을 기록, Llama-3.3-70B, Gemma-3-12B 등 또한 높은 안전성을 보이며 타 모델 대비 선두를 차지
- **(환각률 5%~7%)** GPT-4.1, DeepSeek V3, Qwen3 계열, Gemini-2.5 Pro 등 일반적인 분석·요약 업무에는 활용 가능하나 법·통계 등 정밀 영역에서는 보조 도구로 사용하는 것이 바람직
- **(환각률 8% 이상)** GPT-5 Nano, Claude Sonnet 4, Claude Haiku 4.5, GLM-4.5-AIR의 경우 사실 오류의 가능성이 높은 편이므로, 사실 확인이 필수적인 영역에서는 추가 검증이 필요
- 이외에도 글로벌 AI 싱크탱크 '올어바웃 AI'가 주요 AI 모델의 환각률을 조사한 결과, 의료 정보는 15.6%, 법률정보는 18.7%로 나타났으며, 최신 모델일수록 환각률이 높아진다는 점을 지적[14)]

○ AI 환각의 발생원인

● **(데이터)** LLM의 일반적인 언어 이해·생성 능력과 사실적 지식을 습득하는데 사용된 사전 훈련 데이터, 사용자 지침에 맞게 모델을 정렬하는 단계에서 사용되는 정렬 데이터의 결함이 원인

구 분	주 요 내 용
모방된 거짓	- LLM의 사전 훈련 데이터에 가짜 뉴스나 근거 없는 소문과 같은 잘못된 정보가 포함될 경우, 모델은 이를 사실인 것처럼 암기하고 생성할 가능성이 높음
사회적 편향	- 훈련 데이터에 성별이나 국적과 같은 사회적 편향이 깊이 내재된 경우, 모델이 특정 편향을 증폭시키거나 제공된 문맥과 모순되는 환각을 생성
롱테일 지식*	- 모델이 암기하기 어려운 전문적이거나 드문 지식의 경우, LLM의 사전 훈련 데이터의 범위 내에서 자주 등장하지 않아 환각을 생성하기 쉬운 경향이 존재 * 훈련데이터에서 매우 드물게 등장하지만 그 종류가 방대하여 전체 지식의 상당 부분을 차지하는 전문적이거나 희소한 지식
새로운 사실지식의 도입	- 지도 미세 조정(SFT*) 단계에서 새로운 사실을 통합할 때, LLM이 새로운 지식을 효과적으로 습득하지 못하고 모델 내부에서 환각이 증가될 수 있음 * Supervised Fine-Tuning, 이미 학습된 AI 모델을 특정 목적에 맞게 정렬하기 위해 전문가가 만든 정답 데이터로 추가 훈련시키는 과정, 모델이 원하는 방향으로 정확하게 대답하거나 특정 스타일을 따르도록 조정이 가능
복잡한 지침	- 정렬(Alignment) 데이터에 포함된 사용자의 명령어가 지나치게 복잡한 경우 환각이 발생 가능

● **(훈련방식)** 모델이 언어 구조 및 세계 지식을 습득하고(사전 훈련), 인간의 지침에 맞게 행동하도록 조정되는(정렬) 단계에서 발생하는 훈련방식 내의 근본적인 문제

구 분	주 요 내 용
단방향 예측구조의 한계	- 대부분의 LLM, 특히 GPT 계열 모델은 인과 언어 모델링*을 통해 사전 훈련되는데, 이 과정에서 문맥 간 복잡한 상호의존성을 완전히 반영하지 못해 환각이 발생 * 인과 언어 모델링(causal language modeling)은 GPT 계열 같은 대규모 언어모델(LLM)의 가장 기본적인 학습 방식으로, 앞에 나온 단어들을 보고 다음에 올 단어를 예측하는 단방향적 훈련 방식
노출편향에 의한 오류 증폭	- 모델이 훈련 중에는 정답을 보고 배우지만, 실제 문장 생성 시에는 스스로 만든 이전 단어에 의존해 다음 단어를 예측, 그러나 초기 단어가 잘못 생성될 경우 그 오류가 연쇄적으로 확대
과적합	- SFT단계에서는 모델이 원래 알지 못하는(지식 밖의) 사실이나 정보를 묻는 질문에도 대답하도록 훈련, 이로 인해 자신의 지식 수준을 넘어선 새로운 지식에 맞추고자 과적합이 발생
내부 신념과 출력의 불일치	- RLHF* 과정(Reinforcement Learning from Human Feedback)에서 모델이 진실을 알면서도, 인간 평가자의 선호에 따라 거짓 정보를 출력(아첨(Sycophancy))하여 환각이 발생 * AI 모델에 인간의 평가를 바탕으로 보상을 주어 인간의 선호도에 맞게 학습시키는 기법

● **(추론)** 추론 단계는 LLM이 학습을 마치고 실제 질문에 답을 생성할 때 적용되는 과정으로, 이 단계에서는 디코딩 전략(Decoding*)의 결함이나 모델의 내부적 한계로 인해 환각이 발생

* 다음 토큰(단어)의 확률 분포에서 토큰을 선택하여 문장을 생성해 나가는 과정, 즉 LLM이 실제 답변을 생성하는 과정을 의미

구 분	주 요 내 용
무작위성의 비용	- LLM은 답변의 다양성과 창의성을 높이기 위해 무작위성(랜덤)을 도입한 디코딩 전략을 사용, 이러한 무작위성이 높아질수록 비사실적인 내용을 생성할 위험도 함께 증가
유창성 우선	- 모델이 유창성에 지나치게 집중한 나머지 모델이 부분적으로 생성된 내용에 지나치게 초점을 맞추다 보면 명령어를 망각하고 환각으로 이어질 가능성

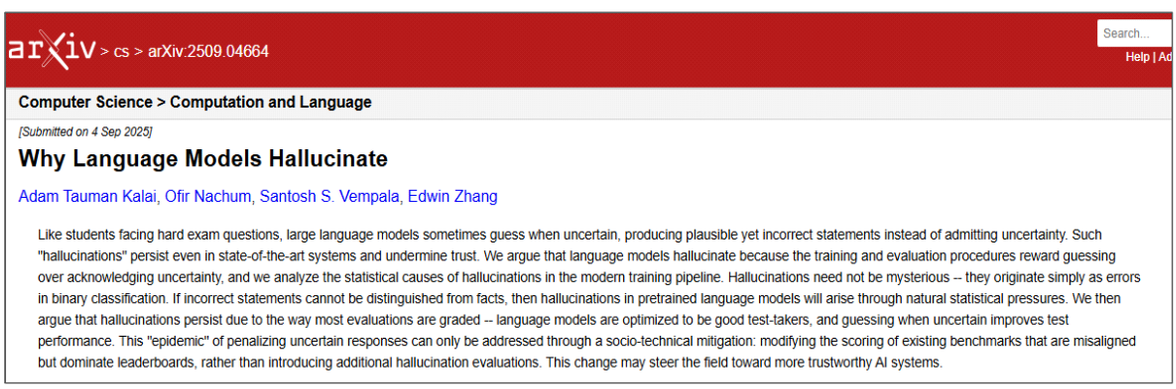
PART III

AI 환각 관련 기업 동향

1 Open AI : AI 환각현상에 대한 평가 기준 개편

- 주요 AI 개발 기업들은 AI 환각이 브랜드의 신뢰도와 서비스의 안정성을 저해하는 요인으로 판단하여, 이를 최소화하기 위한 기술적 및 내부 정책적 조치를 강화하는 중

〈 Open AI가 발표한 'Why Language Models Hallucinate' 〉


arXiv > cs > arXiv:2509.04664

Search...

Help | Ad

Computer Science > Computation and Language

[Submitted on 4 Sep 2025]

Why Language Models Hallucinate

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, Edwin Zhang

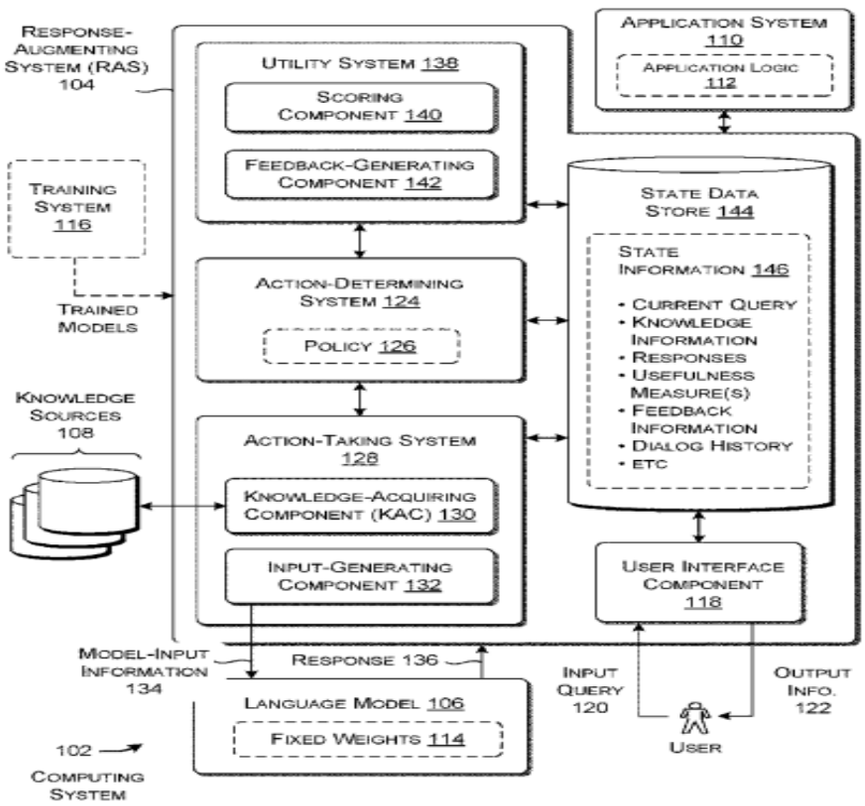
Like students facing hard exam questions, large language models sometimes guess when uncertain, producing plausible yet incorrect statements instead of admitting uncertainty. Such "hallucinations" persist even in state-of-the-art systems and undermine trust. We argue that language models hallucinate because the training and evaluation procedures reward guessing over acknowledging uncertainty, and we analyze the statistical causes of hallucinations in the modern training pipeline. Hallucinations need not be mysterious -- they originate simply as errors in binary classification. If incorrect statements cannot be distinguished from facts, then hallucinations in pretrained language models will arise through natural statistical pressures. We then argue that hallucinations persist due to the way most evaluations are graded -- language models are optimized to be good test-takers, and guessing when uncertain improves test performance. This "epidemic" of penalizing uncertain responses can only be addressed through a socio-technical mitigation: modifying the scoring of existing benchmarks that are misaligned but dominate leaderboards, rather than introducing additional hallucination evaluations. This change may steer the field toward more trustworthy AI systems.

- OpenAI에 따르면 언어모델이 환각을 일으키는 이유는 언어모델의 학습 및 평가구조가 '모른다'고 답하는 것 보다 '추측'을 장려하기 때문에 환각이 발생한다고 설명[15)]
- 가령 기존의 평가 방식은 모델이 정답을 맞히면 점수를 주고, 모른다고 답할 경우에는 0점을 부여하기 때문에 모델은 불확실한 상황에서도 모델은 답을 추측하도록 학습하는 경향이 존재
- 언어모델은 방대한 텍스트 데이터를 기반으로 다음 단어 예측(next-word prediction)을 통해 학습, 맞춤법, 문법처럼 규칙성 있는 패턴은 잘 학습하지만, 무작위적이고 희귀한 정보는 예측하기 어려움
  - ex) 어떤 인물의 생일을 묻는 질문에 모델이 '모른다'고 답하면 0점을 받으나, '9월 10일'이라고 추측하면 낮은 확률로나마 맞을 수 있기 때문에 모델은 정확성보다 자신감 있는 추측을 선택
- 이에 Open AI는 단순히 정확성만 측정하는 대신, '확신에 찬 오답'에는 더 큰 감점을 주고 '모른다'는 겸손한 답변에는 부분점수를 주는 방식으로 평가방식을 수정하도록 권장
- 그러나 일부 질문은 답변 자체가 불가능하거나, 수학적/구조적 한계 때문에 100% 정확도를 달성할 수 없다며 환각의 불가피성을 인정하고 '오류 없는 AI'라는 환상에서 벗어나야 한다고 강조

2 Microsoft : '외부 지식과 피드백을 활용한 언어모델 상호작용' 특허

- 마이크로소프트는 AI 환각 현상이나 잘못된 응답을 줄이거나 방지할 수 있는 기술적 방법에 대한 특허 '외부 지식과 피드백을 활용한 언어모델 상호작용*'을 출원('24.11)[16]

〈 외부 지식과 피드백을 활용한 언어모델 상호작용〉



- 해당 특허는 응답 보강 시스템(Response-augmenting System, RAS*)을 제공하는데 이는 사용자의 질문에 기반해 추가 정보를 수집하고, 응답의 유용성에 대해 검증 가능한 기술을 의미

* RAG(검색증강생성기술)는 AI 모델이 외부 검색결과를 기반으로 응답을 생성하는 기술로 '생성'에 초점을 맞췄다면, RAS는 생성된 응답에 대한 '검증' 작업까지 포함하여 AI모델의 정확성 및 신뢰성을 보다 강화하는 기술

- 또한 RAS는 온라인이나 데이터셋의 '외부 소스'에서 사용자 질문에 대한 답변이 있는지 확인하며, AI가 해당 정보를 답변에 포함하지 않을 경우에는 응답이 부적절하다고 판단하는 기능을 탑재

- 심지어 답변이 부족하거나 의심스럽다고 판단되는 경우, 사용자에게 이를 알리고 피드백을 받을 수 있도록 하며, 이 해결책은 기존 모델의 미세 조정*을 필요로 하지 않는다는 점에서 차이

* 이미 방대한 데이터로 사전 학습(pre-trained)된 대규모 언어 모델(LLM)을 가져와 특정 작업(task)이나 특정 분야(domain)에 더 잘 작동하도록 추가적으로 훈련시키는 과정

- 그러나 AI 응답을 사용자에게 전달하기 전, 응답 내용을 뒷받침하는 실제 데이터 등 AI 응답 내용의 근거 유무에 대한 판단만 가능하기 때문에 거짓 정보 제공을 완전히 방지하는 것은 한계

3 Anthropic : 환각 최소화를 위한 사용자 가이드라인 제공

- 앤트로픽은 Claude 모델의 사용자가 스스로 환각을 최소화할 수 있도록 AI 환각 최소화 가이드라인을 제공하며, 기본적인 환각 최소화 전략과 고급 기술 두 가지로 나누어 안내[17)]
- **(기본적인 환각 최소화 전략)** 불확실성 수용하기, 직접 인용 요청, 인용 검증 3가지로 구성

– **(불확실성 수용하기)** 프롬프트에 모르는 것은 모른다고 말하도록 허용하는 조건을 명시

> 어떤 측면이 불확실하거나 보고서에 필요한 정보가 부족한 경우, 평가하기에 충분한 정보가 없다고 말씀해주세요

– **(사실근거 확보를 위한 직접 인용 요청)** Claude에 작업을 수행하기 전 먼저 원문 그대로의 인용문을 추출하도록 요청(실제 텍스트에 응답의 근거를 두어 환각을 최소화)

> ① GDPR 준수와 가장 관련 있는 정책의 정확한 인용문을 추출하세요. 관련 인용문을 찾을 수 없는 경우, "관련 인용문을 찾을 수 없음"이라고 명시하세요. ② 인용문을 사용하여 이러한 정책 섹션의 준수 여부를 분석하고, 번호로 인용문을 참조하세요. 추출된 인용문에만 기반하여 분석하세요.

– **(인용 검증하기)** Claude가 주장한 내용에 대해 인용문과 출처를 제시하도록 하여 응답을 검증, 인용문을 찾을 수 없는 경우에는 해당 주장에 대해 철회하도록 지시

> ① 해당 제품의 브리프와 시장 보고서의 정보만을 사용하여 사이버보안 제품인 AcmeSecurity Pro에 대한 보도자료를 작성해주세요. ② 작성 후, 보도자료의 각 주장을 검토하세요. 각 주장에 대해 문서에서 직접적으로 지원하는 인용문을 찾으세요. 주장을 지원하는 인용문을 찾을 수 없는 경우, 해당 주장을 보도자료에서 제거하고 제거된 위치를 빈 [] 괄호로 표시하세요.

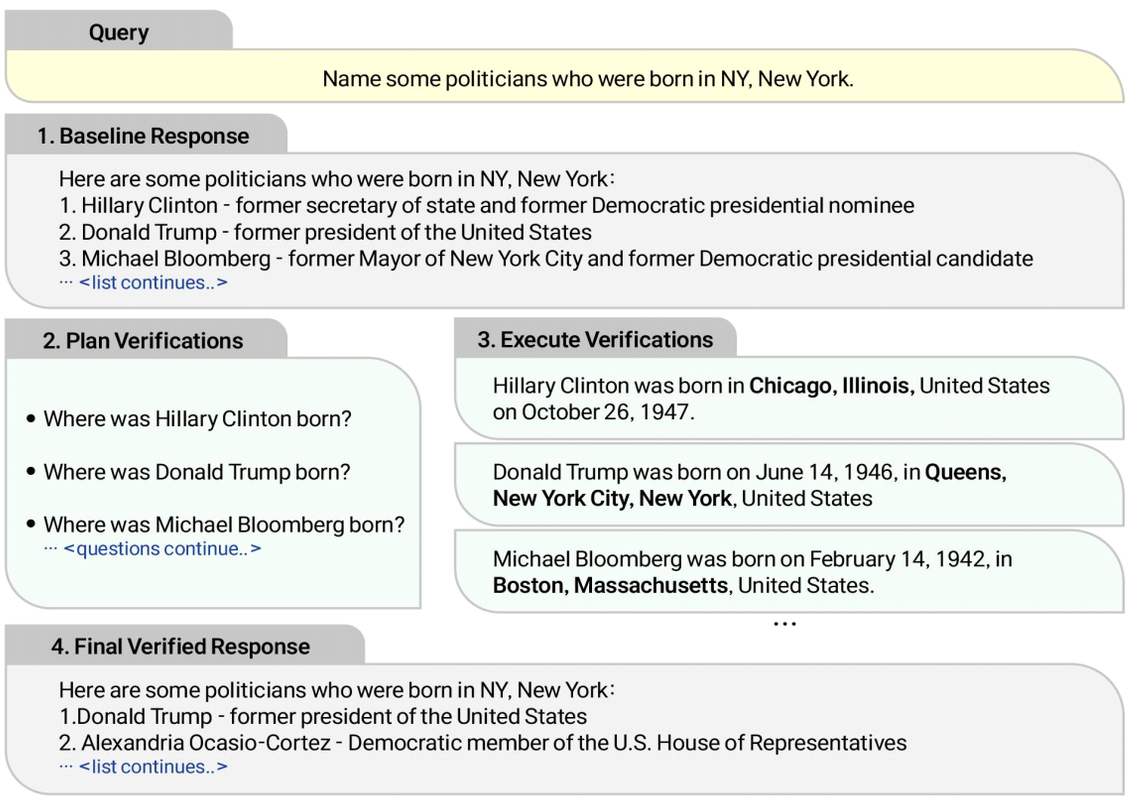
- **(고급기술)** 사고 체인 검증, N번 최적 검증, 반복적 개선, 외부지식 제한 4개 기술에 대해 안내

구 분	주 요 내 용
사고 체인 검증 (Chain-of-thought verification)	– Claude에게 최종 답변을 제시하기 전에 단계별로 추론 과정을 설명하도록 요청하기 – 문제를 단계적으로 분석하도록 지시하여 더욱 정확하고 자세한 결과물 도출을 유도
N번 최적 검증 (Best-of-N verficiation)	– 동일한 질문에 대해 시간차를 두고 2~3번 반복하여 질문하기 – 같은 질문을 여러 번 입력하여 출력이 일관적인지 확인, 출력된 답변 간 내용이 크게 상이한 경우 해당 내용은 신뢰도가 낮다고 판단 가능
반복적 개선 (Iterative refinement)	– Claude의 응답 내용을 후속 질문으로 사용하여 응답 내용을 검토하도록 요청하기 – 예) "당신이 방금 작성한 이전 주장이 이 문서의 내용과 정말 일치하는지 다시 한번 확인하고, 불일치하는 부분이 있다면 수정해주세요"
외부지식 제한 (External knowledge restriction)	– Claude에게 제공된 문서의 정보만 사용하고 일반적인 지식이나 외부 정보는 사용하지 않도록 요청하기(정보 출처를 명확하게 제한하는 지침을 제공)

4 Meta : CoVe 구조 및 HalluLens를 통한 모델 자체의 검증과정 강화

- 메타는 AI 환각을 줄이기 위해 Chain-of-Verification(CoVe) 구조를 개발[18], 이 구조는 모델이 답변을 생성할 때 한 번에 결과를 내지 않고, 스스로의 검증 과정을 거친 후 답변을 생성

〈 CoVe 구조의 작동 방식 〉



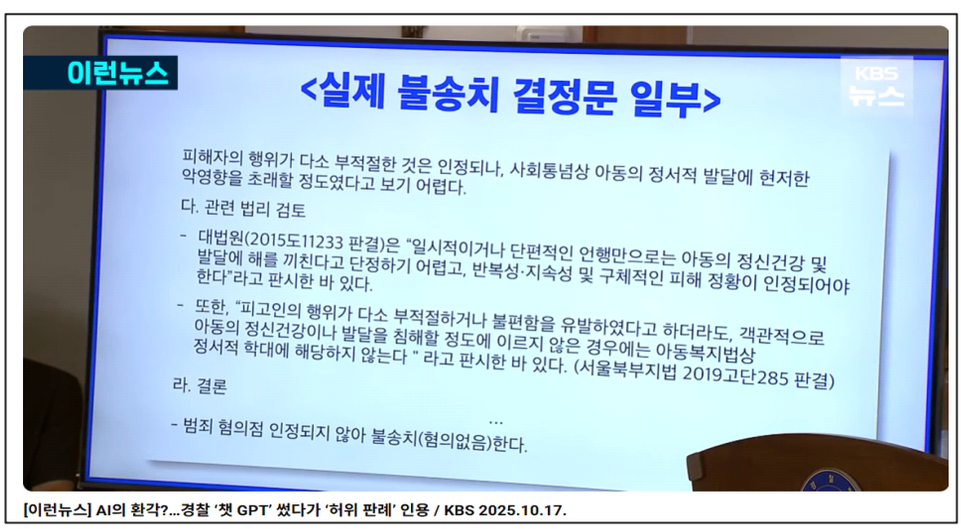
- CoVe는 모델이 먼저 초기 응답을 생성한 뒤, 해당 응답을 사실 확인할 수 있는 검증 질문(verification questions) 생성과 독립적 응답 평가를 거쳐 최종 응답을 도출하도록 설계
- 모델이 응답 생성 과정에서 오류 가능성이 높거나 불확실성이 존재하는 경우에는 응답을 재검토하거나 경고 메시지를 제공하며, 필요 시 사용자 피드백 수집 또한 가능
- 또한 메타는 HalluLens라는 환각 평가용 벤치마크를 개발하여 “내재적(intrinsic)”과 “외재적(extrinsic)” 환각을 구분하고, 모델 응답의 사실성 및 데이터 일관성을 평가
- CoVe와 HalluLens 조합을 통해 장문 생성이나 외부 지식 활용과 같은 다양한 과제에서의 환각 비율의 감소를 확인, 다만 일부 고난도 질문이나 제한된 지식 기반에서는 환각이 여전히 발생
- 이외에도 많은 AI 기업들은 환각 가능성 고지(Disclaimer), 사용자 가이드/베스트 프랙티스 공유, 사용자 인터페이스 기반 가드레일(UI-level Guardrails) 등의 접근 방식을 통해 환각 해소에 노력

Ⅳ PART AI 환각 사례 : 법률, 의료, 과학, 문화

1 법률영역에서의 AI 환각 : AI 환각으로 인한 허위 법리의 인용

- 법률영역에서는 현재 AI가 문서 검토, 리서치 등의 보조적 작업으로 대부분 활용되고 있지만 사소한 오류가 판결·처분결과에 광범위한 영향을 미치며 잘못된 법적 결론으로 연결될 위험
- **(사례)** 경찰청을 대상으로 한 행안위 국정감사에서 경찰이 ChatGPT를 활용하여 아동복지법 위반 사건에 대한 불송치 결정문을 작성, 존재하지 않는 법리를 인용한 사실이 뒤늦게 확인[19)]

〈AI 환각 사례 : 경찰의 챗GPT 활용 불송치 결정문 오류〉



[이런뉴스] AI의 환각?…경찰 '챗 GPT' 썼다가 '허위 판례' 인용 / KBS 2025.10.17.

- 해당 결정문에는 '일시적이거나 단편적인 언행만으로는 아동의 정신 건강 및 발달에 해를 끼친다고 단정하기 어렵다'는 법리가 인용되었으나, 이는 실제 판결문에 없는 내용임을 확인
- 사실관계 및 근거의 정확성이 핵심인 법률 분야에서 AI 환각은 사법적 오판을 쉽게 초래할 뿐만 아니라, 공권력에 대한 국민 신뢰를 심각하게 훼손하는 중대한 위험 요인으로 작용할 가능성

- 이처럼 AI가 날조한 정보를 사용했을 경우 최종적인 법적 책임 소재가 AI 개발사, AI 사용자(경찰/검사 등), 감독자 등 누구에게 귀속되어야 하는지에 대한 법적 논란 발생
- AI 환각은 모델 내부의 복잡한 확률적 예측에서 나오기 때문에 근본 원인을 명확하게 소명하는 데에 어려움이 존재, 이러한 특징은 사법 절차의 투명성을 침해할 가능성이 다분

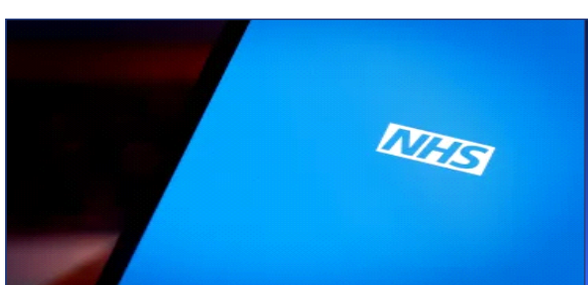
2 의료영역에서의 AI 환각 : 환자 정보에 대한 허위 기록 생성

- 의료 영역에서의 AI 활용은 진단 효율성의 제고와 비용 절감 등 긍정적 효과를 가져오지만, AI 환각이 일어날 경우 환자의 생명과 건강에 직접적인 위협을 가할 수 있는 심각한 문제로 지적
- **(사례)** 영국은 1948년 설립된 'NHS'라는 국영 의료 서비스 시스템을 운영 중이며, 최근 NHS 시스템의 오류로 인해 한 환자에게 당뇨병 환자 대상 정기 안과 검진 초대장이 발송('25.7)
  - 해당 환자는 당뇨병 진단을 받은 적이 없었으며 관련 증상도 부재한 상황으로, 혈액검사 과정에서 간호사와 병력 기록을 조회한 결과, AI가 요약한 환자 기록에서 오류를 확인[20)]
  - 환자는 편도염으로 인해 병원에 방문했으나 기록에는 당뇨병이 진단되어 있었으며 복수의 약물을 복용중이라고 기재, 그러나 실제 증상인 편도염 정보는 누락된 상황
  - 또한 해당 환자의 기록에는 "흉통과 호흡곤란을 호소했으며, 관상동맥질환에 의한 협심증 가능성이 있다"는 허위 정보가 기재, 기록상의 병원 주소 및 우편번호조차 허위 정보로 판별
  - NHS 측은 해당 오류가 '단일 사례의 인간 실수'라고 설명하며, 담당 GP*는 제한적 감독 하에 AI를 활용 중이며, 당시 AI 요약을 검토하던 직원이 초안상태의 원본을 저장한 것이라고 해명

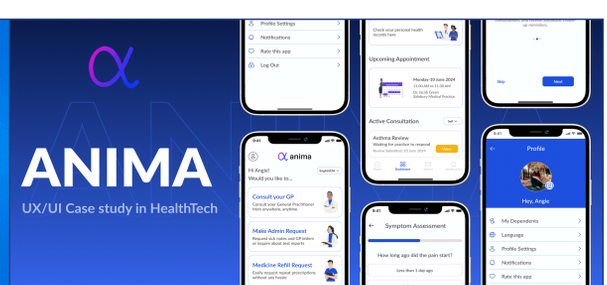
* General Practitioner의 약자로 영국의 1차 진료의를 의미, 우리나라의 동네의원의 일반의, 가정의학과 의사와 유사한 개념

  - 해당 기술을 개발한 애니마 헬스**는 입장 표명을 거부, 이에 NHS 대표는 "애니마는 NHS가 승인한 문서 시스템으로, 문서처리는 사람에 의해 이뤄진다"며 사람의 검토 행위에 대해 강조

** Anima Health : 영국 런던에 본사를 둔 헬스테크 스타트업으로, 환자 요청 제출, 문서 처리, 자동화 분석 기능 등의 제공



영국의 국영 의료 서비스 시스템 'NHS'



NHS에 AI 요약 기능을 제공한 Anima Health

- 위 사례는 AI 환각으로 인해 환자에게 불필요한 안과 검진 및 심리적 불안을 초래하였으며, 실제 당뇨병으로 오인하여 부적절한 처방 및 불필요한 침습적 검사로 이어질 가능성 상존
- 특히 고령자나 문해력이 낮은 환자, 즉 디지털 취약계층의 경우 오류를 인지하지 못하고 AI가 생성한 허위 진단을 그대로 받아들여 심각한 불안을 겪거나 잘못된 생활 습관을 유지할 위험

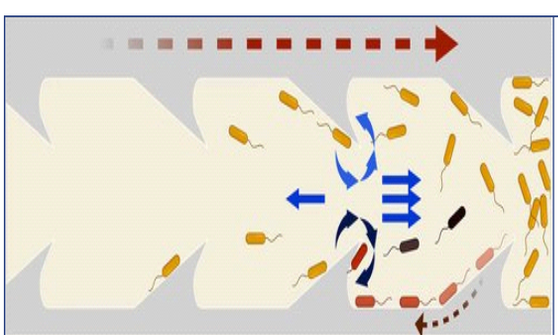
3 과학영역에서의 AI 환각 : 인간 지식을 넘어서는 새로운 아이디어

- 과학 연구분야에서는 AI의 상상력이 인간 지식의 한계를 넘어서는 혁신적인 아이디어를 제공하는 긍정적 효과 또한 발견되며, 현실적 데이터에 기반한 상상력의 촉매제로 재평가
- 미국 연방 AI 연구소장인 에이미 맥거번(Amy McGovern)은 "일반 대중은 AI 환각을 부정적으로만 보지만 실제로는 과학자들에게 새로운 아이디어를 제공하고 있다"고 언급[21)]
- **(사례 1)** 캘리포니아공과대학의 아난드쿠마르 교수의 연구팀은 최근 AI 환각을 활용하여 세균 감염을 크게 줄이는 새로운 카테터*(세균이 부착되기 어려운 삼각형 돌출부가 특징)를 설계[22)]

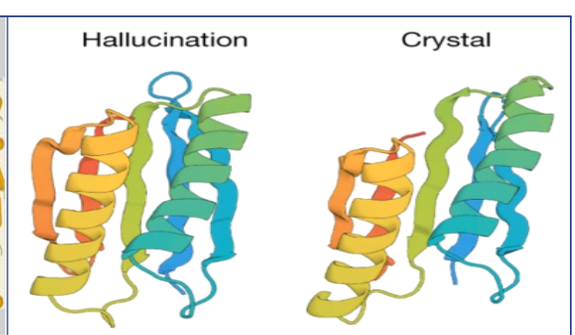
* 카테터 : 인체에 삽입할 때 쓰이는 얇은 의료용 관

- 아난드쿠마르 교수는 "과학분야의 AI는 물리학을 배우고 있다"며 "챗봇과 달리 신뢰할 수 있는 사실에 기반을 두고 있어 매우 정확한 결과를 낼 수 있다"고 설명

- **(사례 2)** 노벨화학상을 수상한 데이비드 베이커 교수의 연구팀은 인간이 모호한 패턴을 의미있는 이미지로 변환하려는 현상인 변상증(Pareidolia)*에서 아이디어를 획득('24.12)[23)]



아난드쿠마르 교수의 새로운 카테터 디자인



Baker Lab의 단백질 설계 연구 예시[24)]

* Pareidolia : 무의미한 자극에서 의미있는 형태나 패턴을 인식하는 현상으로, 실제로는 아무 의미도 없는 모양이나 소리에서 사람 얼굴이나 의미있는 이미지를 보는 착각

- 데이비드 베이커는 "AI 환각이 핵심적인 역할"을 했다며, AI 환각으로 설계된 단백질을 만드는 유전자를 합성해 미생물에 삽입하자 그동안 알려지지 않은 129종의 새로운 단백질이 생성
- 베이커의 연구실은 이 기술로 약 100개의 특허를 획득하고, 20개 이상의 생명 공학 회사를 설립하는 등의 큰 성과를 이루었으며, 최근 더 발전된 기술인 디퓨전(Diffusion) 방식*을 도입

* 디퓨전(Diffusion) 방식 : 노이즈를 제거하면서 원래 데이터(이미지, 소리, 영상 등)를 복원하는 방식

- 달리(DALL-E), 소라(Sora) 등 이미지 생성에 사용되는 디퓨전 방식은 "더 빠르고 성공률이 높다"고 베이커 교수는 평가하며, AI 환각을 통해 새로운 단백질 촉매제를 설계할 것으로 기대

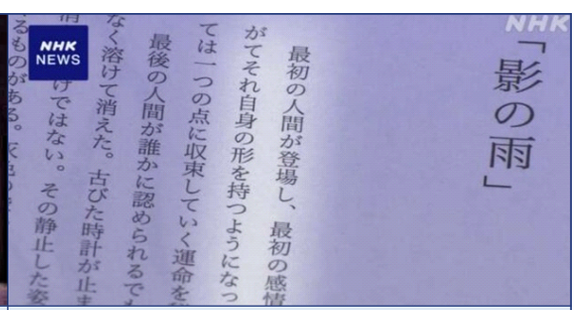
4 문화영역에서의 AI 환각 : 예측 불가능성이 이끄는 초월적 표현

- 과학연구 영역에서는 AI 환각이 새로운 아이디어를 제공하는 상상력의 촉매제로 기능했다면, 문화 영역에서는 AI가 생성하는 예측 불가능한 결과물이 인간 창작의 한계를 확장하는 방향으로 발전
- 다만 AI가 만들어낸 예측 불가능한 결과물에 대해 인간 예술가의 개입정도와 AI의 기여도를 명확히 구분하기 어렵고, 이외에도 데이터 편향, 저작권 등의 문제가 복잡하게 얽혀있는 상황
- **(사례 1)** 미디어 아티스트 레픽 아나돌(Refik Anadol) 그의 대표작 시리즈인 'Machine Hallucinations'를 통해 AI 환각을 시각 예술의 새로운 방법론으로 정립
  - 레픽 아나돌의 작품 중 '희로애락'은 K팝, 뮤직비디오 등 한국과 관련된 다양한 데이터 189만 건을 AI 모델에 학습시켜, AI가 만들어내는 시각적 패턴을 대형 몰입형 설치물로 제작('23.10)
  - 작품 제작 시 AI에게 자신이 원하는 결과물을 세부적으로 명령하는 것이 아니라 AI를 학습시켜 새로운 결과물을 얻는 식으로 제작하여 인간이 상상하기 어려운 독창적 형태와 색채를 확보[25]
- **(사례 2)** 일본의 한 소설 '그림자비*'는 생성형 AI 챗지피티가 95% 집필하고, 나머지 5%를 일본 신인 문학상인 아쿠타가와상을 수상한 작가 구단 리에가 작성하여 화제('25.3)[26]

* 인류가 사라진 뒤 세상에 남겨진 AI가 인간의 기억과 감정의 흔적을 접하며 '감정이란 무엇을 위해 있는가'를 탐구하는 내용의 단편 소설

  - 그림자비의 경우 처음 주제 설정 및 이야기 전개는 모두 AI에게 제안하도록 했으며, 구단 작가는 AI에게 의견을 내거나 방향성을 지시하며 집필을 진행하는 방식으로 약 2주만에 완성
  - 이전에도 구단 리에는 '도쿄도 동정탑'을 발표('24)하며 "해당 작품의 약 5%는 생성형 AI의 문장을 그대로 사용했다"고 밝히며, 이후 잡지사의 요청으로 AI 주필 소설 '그림자비'를 시작

Refic Anadol – 희노애락



AI로 집필된 소설 '그림자비(影の雨)'

- 소설 '그림자비'의 사례는 AI 환각이 문학 창작에서 인간 작가의 관습적인 서사 구조를 해체하고 새롭고 실험적인 전개를 촉발할 수 있음을 시사하며, 신선한 문학적 경험을 제공

PART V

AI 환각 관련 해외 정책 동향

1 EU : AI Act 이행 지원을 위한 실무적 권고 지침 마련

- 많은 국가가 직접적으로 환각이라는 용어로 규제하지는 않으나, 생성형 AI가 만드는 부정확·허위 콘텐츠를 억제하는 조치를 통해 사실상 환각 감소를 목표로 하는 규제를 도입 중
- 핵심 수단은 ① AI 생성물 표시 및 라벨링, ② 훈련 데이터·출처 투명성 확보, ③ 개발 단계의 검증·테스트(리스크 평가 등), ④ 플랫폼·서비스 사업자에게 부여되는 필터링·수정·보고 의무
- EU는 AI Act를 통해 생성형 AI의 투명성·추적성·사전 검증을 법제화하여, AI가 만든 콘텐츠에 대한 표시 의무와 고위험 AI의 사실성 검증·로그 보관을 요구
  - AI Act 및 관련 가이드라인(General Purpose AI 지침 및 투명성 실천 강령* 등)을 통해 GPAI(범용AI)에 대한 적합성 평가·사전 리스크 관리·사후 추적성 요구를 구체화[27)]

* AI법 제 50조(투명성 의무)의 이행을 지원하기 위해 설계된 실무 지침으로, 법으로 강제되지는 않으나, 법 준수를 돕는 자발적 권고 지침, 특히 딥페이크나 조작된 콘텐츠, "사람을 속일 수 있는 합성 콘텐츠"에 대한 투명성 확보에 목표

〈 Code of Practice on transparency of AI-generated content 내용 〉

구 분	주요내용
개발자의 의무 (Providers)	- 생성형 AI 시스템의 출력물(텍스트, 이미지, 음성 등)은 머신 판독 가능(machine-readable), 즉 기계 혹은 컴퓨터가 사람의 개입 없이 읽어들일 수 있는 데이터 포맷으로 표시할 필요 - 사용되는 기술 솔루션은 가능하면 효과적이면서도 상호운용성(interoperable)이 있어야 하고, 견고하고 신뢰할 수 있어야 함(robust & reliable) - 구현 시 콘텐츠 유형(텍스트, 음성, 비디오 등)의 특성, 기술 한계, 비용 등을 고려하여 설계
공개/배포자의 의무 (Deployers)	- 딥페이크(사람, 사물, 장소, 사건 등을 사실처럼 보이게 조작한 이미지·오디오·비디오)에 대해서는, AI 생성 또는 조작되었다는 것을 공개할 것 - 공공의 이익과 관련된 텍스트(예: 뉴스, 공공 정보) 중 AI가 생성 혹은 조작한 것은, 사람이 검토한 경우를 제외하고 AI가 만들었다고 명시

- 이러한 사전 검증·추적성 강화 등의 방안은 AI 응답내용의 출처·근거를 확인하여 환각의 발생 빈도를 줄이는 데 유리, 그러나 전문영역의 '정답 불명확성**' 문제는 여전히 한계

** 전문 지식, 최신 연구 결과, 의학 진단 등 전문가 사이에서도 의견이 다른 영역이거나, 정답이 하나로 정해지지 않은 영역에 대한 문제

2 미국 : 혁신을 최우선하기 위한 정부 개입의 최소화

- 바이든 행정부 집권 당시 미국 국립표준기술원의 AI 위험 관리 프레임워크(AI RMF)를 통해 AI 환각이나 허위 정보 생성에 대한 위험을 식별하고 이를 관리하기 위한 지침을 제공('24)
  - NIST는 환각을 생성형 AI의 자연스러운 결과이자 심각한 신뢰성 위험으로도 정의하며, 환각(Confabulation*)에 대한 AI RMF의 네 가지 핵심 기능의 걸친 체계적인 대응을 요구[28]

* Hallucination은 AI가 내부적으로 사실과 다른 정보를 생성하는 기술적 오류를 의미, 한편 Confabulation(허구/날조)의 경우 기술적 오류보다는 사용자를 속일 수 있는 결과로 이어지는 '위험'의 측면을 강조(본 보고서에서는 '환각'으로 용어를 통일)

〈 AI RMF : 생성형 AI 프로파일에서의 환각(Confabulation) 관련 내용 〉

RMF 기능	핵심목표	주요 요구사항
매핑 (Map)	위험 식별 및 한계 정의	- **(위험 식별 및 문서화)** 환각의 발생범위(AI시스템이 무엇을 모르는지)에 대해 시스템의 사실성·정확성 한계를 명확하게 식별하고 문서화**
측정 (Maeasure)	정량적 평가 및 추적	- **(객관적 평가)** 경험적으로 검증된 방법으로 모델의 역량을 평가하고, 전문 테스트 환경을 활용하여 신뢰성을 평가, 또한 배포 전후 GAI 출력의 출처 검토·확인 등
관리 (Manage)	위험 완화 및 제어	- **(운영적 통제 및 배포 후 관리)** 조직의 위험 허용 범위를 벗어나는 모델은 재훈련 또는 폐기, 배포 후에는 환각, 사이버 위험 등에 대비하여 지속적으로 모니터링
거버넌스 (Govern)	거버넌스 및 정책 수립	- **(배포 통제)** GAI 출시 전, 성능과 위험 측정 결과를 반영한 최소 안전 기준을 설정하고,이를 충족해야만 배포를 승인하는 go-no-go 정책을 확립 - **(지속적 개선 및 투명성)** 위험 측정 프로세스를 정기적으로 평가하고 업데이트하는 정책을 확립, 또한 사용자에게 GAI로 야기되는 위험을 미리 고지

** 해당내용(Map 2.2)의 경우, Human- AI Configuration(AI에 대한 과도한 신뢰 등 AI를 사용하는 인간의 심리적 반응으로 발생하는 모든 문제를 의미) 위험에 대한 대응사항으로 명시되어 있으나, 간접적인 환각의 대응조치로 간주하여 포함

- 그러나 트럼프 행정부 2기는 'Executive Order 14179 Removing Barriers to American Leadership in Artificial Intelligence(AI 분야에서의 선도를 위한 장벽 제거)*'를 발효('25.1)

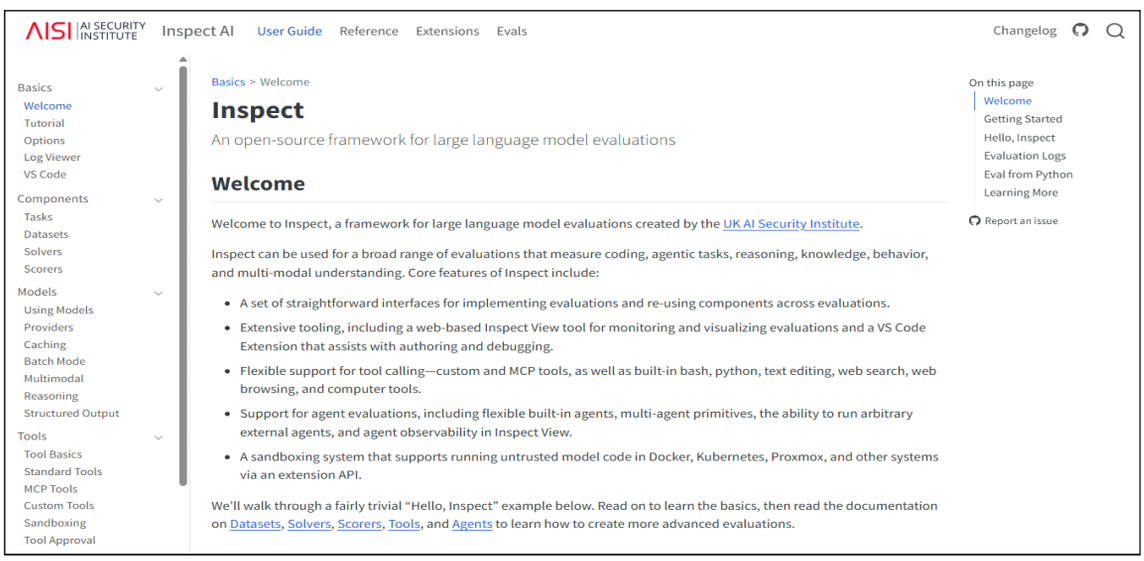
* 해당 명령은 미국의 AI 혁신 리더십을 강화하기 위해 기존 일부 정책을 철회하고, 정부 정책 장벽을 제거하는 것이 목표

  - 해당 행정명령은 AI의 신뢰성과 통제를 언급하고 있으나, AI가 만들어낸 정보에 대한 객관성 및 정확성보다는 '이념적 편향(ideological bias) 제거'에 초점
  - 또한 해당 행정명령은 AI 개발에 대한 안전성·신뢰성을 중요시했던 이전 바이든 행정부의 행정명령(Ex) EO 14110)을 재검토하고, 일부 정책을 철회하거나 수정하도록 지시

- 현재 미국은 이전 바이든 정부와 달리 기술 혁신을 저해하는 규제를 최소화하면서도, AI 규제의 초점을 환각과 같은 기술적 위험보다는 편향성 제거를 통한 AI의 객관성·공정성 확보에 집중

3 영국 : 법적 강제 없는 유연한 AI 안전 표준 주도

- 영국의 경우 AI 환각 위험에 대해 유연한 규제 환경을 갖춤으로써, 기술 자체의 위험을 정량적으로 측정하고, 그 결과를 바탕으로 AI 개발사들의 자발적인 안전 조치를 유도하는데 초점
- 영국의 AI 환각 대응 전략은 기술적 통제 및 측정과 정책적 유연성이라는 두 축으로 나눌 수 있으며 기술적 통제 및 측정의 경우 UK AI 보안연구소를 중심으로 추진 중
- UK AI 보안연구소는 AI의 안전성 측정을 위한 오픈소스 평가도구 'Inspect AI'를 개발[29], Scorers 기능을 통해 모델의 정확성과 일관성을 정량적으로 측정하여 환각 발생 비율을 산출



〈 InspectAI의 3대 구성요소 〉

구성요소	역할
Datasets (데이터셋)	- 평가에 사용할 모든 입력 프롬프트, 질문, 명령어 및 이에 대한 기대되는 정답이나 채점 기준을 정의 - 모델에게 물어볼 데이터의 품질, 다양성, 그리고 환각을 유발할 수 있는 취약성 질문 등을 체계적으로 포함
Solvers (실행(모듈))	- Datasets의 입력을 받아 모델에 전달하고 응답을 얻어내는 과정을 제어하는 실행 논리 모듈 - 단순 API 호출을 넘어, 복잡한 멀티-턴 대화 시퀀스나 특정 역할극 지시 등 레드팀 운영 시나리오를 구현하여 모델의 안전 장벽을 우회하고 환각을 유발하는 방식을 실행
Scorers (평가(모듈))	- Solvers를 통해 모델로부터 얻은 응답(출력)을 평가 기준에 따라 점수화하고 정량적 지표로 변환 - 모델 응답이 외부의 신뢰할 수 있는 사실(Grounding)과 일치하는지 검증하여 환각 발생 비율을 계산, 이를 통해 모델의 견고성과 신뢰도를 객관적인 수치로 제공

- 영국은 EU와 달리 AI 신뢰성, 안전 등에 대해서 법적으로 강제하지 않으며, 또한 AI 환각 문제를 윤리적 문제가 아닌 측정 및 통제 가능한 기술적 결함으로 정의하고 대응하는 기조

4 싱가포르 : AI 신뢰성 관련 기술 프레임워크 및 툴킷 개발

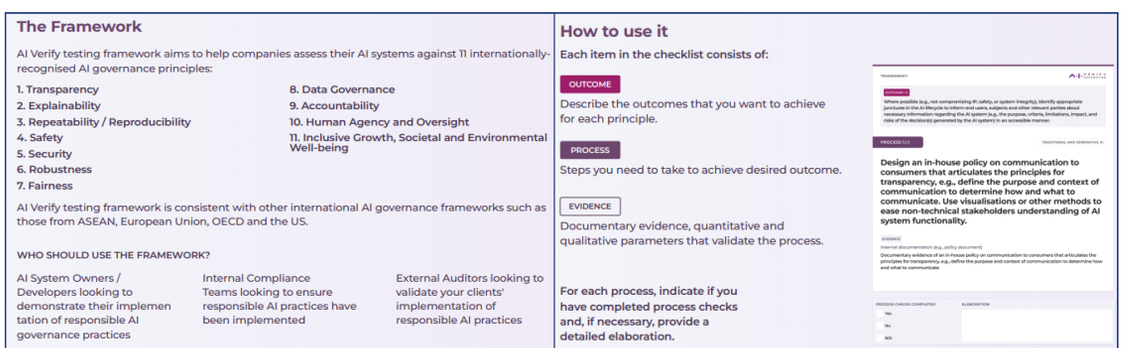
- 싱가포르의 대응전략은 정보통신미디어발전청(IMDA)를 중심으로 이루어지며, 환각 등의 AI 신뢰성 문제에 대한 법적 강제보다는 '신뢰 구축' 및 '실용적인 기술 프레임워크'에 중점
- 싱가포르는 AI 모델의 책임 있는 사용을 입증하기 위한 기술적 거버넌스 프레임워크 및 소프트웨어 툴킷인 AI Verify 테스트 프레임워크*를 개발하며, 11가지 AI 거버넌스 원칙을 제시[30)]

* AI 모델의 안정성, 공정성, 설명가능성, 유해성 등을 점검하는 프레임워크

〈 11가지 AI 거버넌스 원칙 내용 〉

순번	원칙	주요내용
1	투명성	AI 시스템이 어떻게 사용되는지, 어떤 목적으로 의사결정에 활용되는지 사용자에게 알릴 수 있어야 함
2	설명가능성	AI가 내린 결정이나 예측의 근거를 설명할 수 있어야 함
3	반복성/재현성	동일한 입력을 주었을 때 AI 모델이 일관된 결과를 낼 수 있어야 함
4	안전	AI는 오작동, 잘못된 예측 등으로 인해 사용자나 사회에 해를 끼치지 않도록 설계되어야 함
5	보안	AI 시스템, 데이터 및 관련 인프라를 무단 접근, 공개, 수정 등으로부터 보호하여야 함
6	견고성	AI 시스템은 입력 변화, 환경 변화, 악의적 조작에도 안정적으로 작동할 수 있어야 함
7	공정성	AI가 어느 특정 그룹(예: 인종, 성별, 연령 등)에 불공정한 편향(bias)을 가지지 않도록 해야 함
8	데이터 거버넌스	데이터 출처, 품질, 보존, 데이터 변경 이력 등 데이터의 출처(lineage)를 명확히 관리해야 함
9	책임성	잘못된 AI 행동에 대해 책임을 지고 대응할 수 있어야 하며, 책임 주체를 명확히 정의해야 함
10	인간주체성 및 감독	AI가 의사결정을 보조하더라도 사람이 최종 통제권을 가질 수 있어야 함
11	포용적 성장, 사회 및 환경적 웰빙	AI의 발전이 단순한 기업 성장이 아니라 모든 계층의 사회적 이익으로 이어지도록 고려해야 함

- 광범위한 AI 위험을 '테스트 기반'으로 검증하려는 최초의 국가라는 점에서 정책적 의미가 있으며, 환각 해소를 위한 직접적 규제는 아니나 검증 체계 마련을 통한 AI 부작용 피해를 최소화



AI Verify 프레임워크 설명 | 툴킷 사용 방법

- 이뿐만 아니라 싱가포르는 AI 및 데이터 거버넌스 전략을 뒷받침하기 위해 디지털 신뢰 센터(Digital Trust Centre)을 설립('21~)하는 등, 글로벌 AI 신뢰 허브로서의 입지를 강화[31)]

PART VI 정책적 시사점 및 대응 전략

○ AI 환각의 문제 정의 및 현황

- AI 환각은 생성형 AI가 사실과 다른 정보를 진짜처럼 제시하는 현상을 의미하며, 최근 AI 활용이 증가하면서 환각은 단순 오류를 넘어 사회적·경제적·법적 리스크를 야기하는 주요 문제로 부상
- 의료 영역에서는 AI가 존재하지 않는 진단이나 약물정보를 제공하여 판단에 영향을 주거나, 금융 분야에서는 잘못된 투자 정보로 인해 경제적 피해가 발생하는 등 AI자체의 신뢰성이 훼손
- AI 환각은 모델 구조, 학습 데이터 품질, 추론과정의 불확실성 등 기술적 요인에 의해 발생하나 사용자의 검증 능력과 상호작용에 따라 사회적 피해가 증폭될 수 있음
- 반면 AI 환각 현상은 위험을 내포하면서도 새로운 아이디어를 제공하거나 예술 창작, 연구가설 탐색 등의 과정에서 기술 발전 및 창의성의 동력으로 작용 가능
- 따라서 AI 환각에 대한 대응은 단순 기술적 문제 해결에 국한되지 않고, 위험을 관리하면서도 잠재적 가치를 활용할 수 있는 사회적·정책적 차원에서의 균형잡힌 포괄적 접근이 필요한 상황

○ '환각' 개념의 재정립: 결함(Defect)이 아닌 기능(Function)으로의 전환

- 현재 'AI 환각(AI Hallucination)'이라는 용어는 오류·오작동의 뉘앙스를 강하게 내포하여 정책·사회적 담론에서 '필연적이고 위험한 결함'으로 받아들여지고 있음
- 그러나 생성형 AI의 작동 원리를 고려하면 AI 환각은 확률적 생성 매커니즘이 만들어내는 자연스러운 부작용이자 특징에 가까움, 즉 모델은 새로운 조합·추론적 비약 등이 불가피하게 발생
- 따라서 오류 중심의 규제 접근 → 생성 중심의 관리 접근, '허위 정보 생성에 대한 비난적 프레임' → '어떤 조건에서 어떤 유형의 환각이 발생하는가'를 이해하는 기술적 프레임으로의 관점 전환이 필요
- 이는 단순히 환각현상을 규제하거나, 관리하는데에 그치지 않고 '환각'이 지닌 부정적 이미지에서 벗어나 환각의 기술적·기능적 특성을 정확히 반영하도록 중립적 용어로써 재해석 할 필요
- 'AI 환각'이라는 단어는 이미 산업·학계에서 통용되는 용어지만, 관점 전환을 위해 그 개념을

'추론적 편차(Inferential Deviation)'와 같은 중립적·기능적 용어로 재해석할 것을 제안

- 그러나 환각의 위험을 고려할 때 정책·규제·감독 체계 내에서 명확히 식별될 필요가 있으므로 오정보 출력, 창의적 생성 등의 용어로 부정적, 긍정적 효과에 따른 새로운 하위개념으로 분리

- 이러한 환각 개념 정의 및 유형의 분리는 단순한 개념 정리 차원을 넘어 위험관리 기반 접근과 창의적 활용 전략을 동시에 가능하게 하는 정책적 기반으로서 작용하는 첫 걸음

〈 AI 환각의 개념 정의 및 유형체계(안) 〉

용 어	의 미	비 고
AI 환각 (AI Hallucination)	AI가 확률적 과정에서 사실과 불일치 하는 출력을 내는 현상	기존의 학계·산업계 통용 용어
추론적 편차 (Inferential Deviation)	AI가 학습된 패턴을 바탕으로 새로운 조합이나 예측을 만들면서 나타나는 확률적 편차로, 생성형 AI에서 나타나는 기술적 특징	환각을 설명하는 가치 중립적 개념
오정보 출력 (Factual Error Output)	사회적·경제적 피해를 유발할 가능성이 있는 사실 왜곡 출력 현상	가치판단 용어(부정), 규제 및 위험 관리 대상
창의적 생성 (Creative Production)	AI 모델이 새로운 아이디어, 조합 등을 만들어내는 출력 방식	가치판단 용어(긍정), 혁신 및 진흥 대상

○ 위험 관리 중심의 균형적 대응 전략 : CURE(Control – Utilize – Regulate – Evaluate)

- 국제적으로 AI 환각에 대한 대응은 주로 오정보 방지 중심으로 설계되어 왔으며, 글로벌 빅테크 기업 역시 검증 절차, 지식 기반 보강(RAG)등 정확성 제고 중심의 기술적 수단을 고도화 중

- 이러한 접근은 부정적인 환각을 억제하는데 일정한 효과가 있지만, 동시에 생성형 AI가 갖는 창의적·확장적 특성까지 규제 프레임에 포섭하는 경향, 혁신 가능성을 함께 제한하는 구조적 한계

- EU의 AI Act는 AI의 고위험 영역에 대한 의무 부과, 미국은 자율 기반의 기술적 안전성 확보에 초점을 맞추는 등 각국의 대응은 '위험 관리'에 치중, 환각의 생산적 활용에 대한 전략적 관점이 부족

- 이러한 국제적 흐름 속에서, 우리는 AI 환각을 위험과 기회가 공존하는 양면적인 현상으로 보고, 이를 균형있게 다루는 정책 프레임워크 CURE Framework를 제안

위험 관리 중심의 균형적 대응 전략 : CURE Framework	
Control : 오정보 출력에 대한 정밀 통제	Utilize : 창의적 활용을 위한 리터러시 강화
Regulate : 법·제도적 관리 체계 정립	Evaluate : 지속적 평가와 모니터링

- 해당 프레임워크는 오정보 출력 등의 부정적 환각은 통제(Control)하여 위해를 최소화하되, 창의적 생성류의 긍정적 환각은 활용(Utilize)할 수 있도록 도움으로써 혁신 역량을 극대화
- 또한 변화하는 생성형 AI 환경에 맞추어 법·제도적 기반(Regulate)을 정교하게 설계하며, 실제 서비스 맥락에서 환각 발생 경향을 지속적으로 평가(Evaluate)하여 기술 및 정책을 보완

● **(Control : 오정보 출력에 대한 정밀 통제)** 사실과 불일치한 출력, 오정보성 환각은 사회적·경제적 위해를 초래할 수 있으므로 선제적 통제 장치가 필수적

- 신뢰 가능한 외부 지식 연계(RAG·지식기반 검증), 출처 투명성 확보, 사실 검증 절차의 자동화 등 기술적·운영적 수단을 체계적으로 마련하여 오류 확산 가능성을 최소화

● **(Utilize : 창의적 활용을 위한 리터러시 강화)** 오정보 출력과 달리, 새로운 조합과 독창적 아이디어를 생성하는 창의적 측면에서의 환각은 AI의 혁신 동력으로 발전될 가능성이 존재

- AI 환각을 적극적으로 활용하여 인간의 아이디어 발굴·문제 재구성·창작 과정의 효율을 높이면서도, 부정적 환각을 줄이기 위한 '창의적 프롬프트' 가이드라인 개발·교육

〈 창의적 프롬프트 가이드라인(안)* 〉

순번	원 칙	주 요 내 용
1	창의적 환각을 위한 모드 설정 (불확실성 유도)	- '팩트 체크'가 아닌 '개념 조합'에 집중하도록 허용하는 조건을 명시하기 ex) ① "이 답변은 사실적 정확도를 목표로 하지 않습니다", ② "현재 지식에 기반한 현실적 제약을 무시하고, 가장 비현실적이고 독창적인 아이디어를 3가지 제시해주세요"
2	창의적 제약 조건 명시 (이질적 요소의 결합)	- 기본적인 AI의 사고를 전환시키기 위해 이질적 요소를 결합하도록 요청하기 ex) ① "이 문제를 해결하기 위해 전혀 관련없는 분야의 개념 2가지(고전미술, 생물학)를 강제로 연결시키세요" ② "답변 시, 셰익스피어의 비극 문체와 SF 영화 시나리오의 문체를 동시에 사용하세요" ③ "인용문을 사용할 경우, 실존하지 않는 가상의 철학자나 미래의 역사책에서 가져온 것처럼 만드세요"
3	AI 응답결과의 아이디어화 (비판적 사고 유도)	- AI모델의 응답 결과를 최종답안이 아닌 아이디어의 원천으로 활용하도록 만들기 ex) ① "생성된 아이디어 중 가장 비현실적인 아이디어 1가지를 선택하고, 이것이 현실에서 가능하게 할 수 있는 최소한의 조건 3가지를 역으로 분석해 주세요"(AI응답에 대한 비판적 사고 유도) ② "'이것은 아이디어입니다. 실제 적용 전에는 반드시 관련 전문가의 타당성 검토가 필요합니다.'라는 문구를 답변 하단에 포함하세요"(AI 모델에 '사실'이 아닌 '아이디어'임을 강조)

※ 해당 가이드라인은 AI 응답결과의 유창성, 독창성, 정교성, 유연성 등 객관적인 창의성 지표 개발을 통한 검증 연구가 선행될 필요

● **(Regulate : 법·제도적 관리 체계 정립)** AI 환각이 초래하는 위해성은 기술적 문제를 넘어 사회적·경제적 피해로 확대될 가능성, 사고 발생 시, 책임 소재를 명확히 하여 신뢰성을 확보

- AI 활용 분야별 위험도에 따른 차등형 규제제도를 마련하고, 특히 고위험 영역(의료·법률·공공 등)에 대해서는 보다 강화된 검증 체계(AI 안전 검증 위원회 등)를 도입함으로써 AI의 안전성을 담보

〈 영역별 환각 위험도 및 정책 방향 요약 〉

영역	특징	환각 위험 수준	정책 방향	규제 필요성
의료 진단·판독·진료	환자 안전 직결	초고위험(🔴)	검증·감독·책임성 강화	매우 높음
공공행정·사법 판단	사회적 피해	고위험(🟠)	감독·증거 기반 사용	높음
교육·언론	정보 신뢰성 중요	보통(🟡)	출처 표시·팩트체크	보통
연구개발·신약 탐색	창의적 발상 필요	저위험(🟢)	환각의 창의적 활용 장려	낮음
예술·콘텐츠 제작	창작성이 핵심	초저위험(🟢)	환각 적극 활용 가능	매우 낮음

- **(Evaluate : 지속적 평가와 모니터링)** AI의 환각 발생률은 모델 구조, 학습 데이터, 운영 맥락에 따라 지속적으로 변동하므로 도입-운영 -업데이트 전 주기를 아우르는 연속적 평가체계가 필요

  - 출력 오류율 모니터링, 인간 검증자(Human-in-the-loop) 평가, 실제 서비스 환경 기반의 정성·정량 지표 관리 등을 통해 환각 발생 추세를 수집·분석, 그 결과를 기술 개선과 정책 조정에 반영

○ 종합 시사점

- AI 환각은 기술 발전 속도에 비해 대응 체계가 충분히 정비되지 않아, 통제·활용·규범화가 요구되는 복합적 위험요소로, 이에 대한 사회적 논의와 정책적 준비가 병행되어야 할 시급한 과제
- 특히 고령층의 경우 환각 현상 자체에 대한 인지 역량이 낮으며, 높은 신뢰도가 요구되는 의료, 행정영역에서의 환각(오정보 등)로 인한 피해는 생명 또는 재산에 직접적인 위해가 될 수 있음
- 반면 생성형 AI의 확산으로 AI 환각 문제는 다양한 분야에서의 위험과 혼선을 초래함과 동시에 새로운 아이디어의 발굴 등 창의적 활용을 통한 '생산성 향상의 기회'로도 인식
- 이러한 특성을 고려할 때, AI 환각을 단순히 제거·억제의 대상이 아닌, AI 모델 특성·AI 활용 분야의 위험도, 사용 맥락에 따라 달리 관리해야 하는 현상으로 인식하는 것이 중요
- 본 보고서에서 제시한 방향성은 완결된 해답이 아닌 정부·산업·이용자 모두가 활용·참고할 수 있는 정책적 출발점으로, 실제 서비스·산업별 적용과정에서의 다양한 사례 축적이 필수적
- 나아가 향후 AI 환각에 대한 국제 표준 정립, 고위험군 분야 대상의 AI 안전 제도 마련, 이용자·조직 차원의 교육·리터러시 확산 등을 거치며 지속적으로 보완될 수밖에 없는 한계가 존재
- 궁극적으로 AI 환각 대응은 'AI 기본사회'의 신뢰 기반을 형성하는 핵심적 과제로 인식하고, 사회적 합의와 이행 경험을 축적해가는 지속적·점진적 정책 과정으로 다뤄져야 할 필요

참 / 고 / 자 / 료

1) https://m.news.nate.com/view/20250213n26882?mid=m04&list=recent&cpcd=
2) https://www.lawtimes.co.kr/news/197090
3) https://apnews.com/article/ai-artificial-intelligence-health-business-90020cdf5fa16c79ca2e5b6c4c9bbb14
4) https://aimatters.co.kr/news-report/ai-news/11223/
5) https://openai.com/ko-KR/index/why-language-models-hallucinate/
6) https://share.google/SyUDdt5pCc6w2Unf8
7) https://news.microsoft.com/source/features/company-news/why-ai-sometimes-gets-it-wrong-and-big-strides-to-address-it
8) https://www.ibm.com/think/topics/ai-hallucinations
9) https://www.fnnews.com/news/202303151037473909
10) LEI HUAN 외 5인(2023.11), A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions
11) Survey of Hallucination in Natural Language Generation
12) https://link.springer.com/article/10.1007/s10579-025-09864-x
13) 'All Hallucinations are Not Bad. Acknowledging Gen AI's Constraints and Benefits'
14) https://www.allaboutai.com/resources/ai-statistics/ai-hallucinations/
15) https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf
16) https://aimatters.co.kr/news-report/ai-news/8317/
17) https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-hallucinations
18) https://aclanthology.org/2024.findings-acl.212/
19) https://www.youtube.com/watch?v=N700aRVBRVw
20) https://www.fortunekorea.co.kr/news/articleView.html?idxno=49097
21) https://fortune.com/2024/12/24/ai-hallucinations-good-for-research-science-inventions-discoveries
22) https://www.caltech.edu/about/news/aided-by-ai-new-catheter-design-prevents-bacterial-infections
23) https://m.dongascience.com/en/news/69171
24) https://www.bakerlab.org/2021/12/02/deep-learning-protein-design/
25) https://www.hankyung.com/article/2023101078341
26) https://www.jk-daily.co.kr/news/articleView.html?idxno=24395
27) https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai?utm_source=chatgpt.com#ecl-inpage-Signatories-of-the-AI-Pact
28) https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence?utm_source=chatgpt.com
29) AI Security Institute, UK. 2024. "Inspect AI: Framework for Large Language Model Evaluations." https://github.com/UKGovernmentBEIS/inspect_ai.
30) https://aiverifyfoundation.sg/what-is-ai-verify/
31) https://www.ntu.edu.sg/dtc