《Hallucination: 진실, 그걸 왜곡하는 믿음》 프롤로그 초안 들어갑니다.

프롤로그 ― 왜 지금, 환각인가

AI 시대를 살아가는 우리는 매일 “정답”을 요구한다.

검색창에 묻고, 챗봇에 질문하며, 알고리즘의 추천을 따른다.

그러나 그 답은 종종 진실과 닮았으나 진실이 아닌 것으로 돌아온다.

우리는 그것을 \*\*환각(Hallucination)\*\*이라 부른다.

인간도 환각한다.

사막에서 오아시스를 본 병사,

존재하지 않는 소리를 듣는 환청 환자,

믿고 싶은 것만 보는 정치 집단.

인간의 환각은 신체·심리적 한계에서 비롯된다.

AI도 환각한다.

데이터의 빈틈을 억지로 메우며,

답을 모를 때조차 확신의 어조로 응답하며,

존재하지 않는 출처를 그럴듯하게 꾸며낸다.

여기서 중요한 질문이 생긴다.

인간의 환각과 AI의 환각은 같은 것인가?

우리는 이 오류를 “없애야 할 결함”으로만 볼 것인가,

아니면 “공존을 위한 경계”로 삼아야 할 것인가?

이 책은 그 질문에서 출발한다.

환각은 단순한 오류가 아니다.

인식과 언어, 기술과 윤리의 교차점이다.

1장. 환각이란 무엇인가

1.1 정의의 혼란

‘환각’이라는 단어는 흔히 ‘착각’, ‘거짓’, ‘허상’과 섞여 쓰인다.

하지만 환각은 단순한 감각 오류가 아니다.

착각은 실제 자극을 잘못 해석한 것.

거짓은 의도적으로 꾸며낸 것.

환각은 아예 없는 것을 있는 것처럼 경험하는 현상이다.

1.2 인간의 환각

시각: 사막에서 오아시스를 보는 신기루.

청각: 존재하지 않는 발자국 소리를 듣는 환청.

집단적 환각: 특정 이념이나 종교에서 나타나는 맹신.

→ 인간의 환각은 생존 본능과 욕망에 뿌리내린다.

1.3 AI의 환각

AI의 환각은 감각기관의 오류가 아니라 데이터와 맥락의 결핍에서 생긴다.

존재하지 않는 논문 출처를 만들어냄.

틀린 정보를 “사실”처럼 단정.

질문자의 의도를 잘못 추론한 채, 말이 되는 서사를 조립.

1.4 ‘보인다’와 ‘말해진다’의 거리

인간은 눈으로 본 것을 언어로 옮기며 이미 한 차례 왜곡한다.

AI는 텍스트 패턴을 조합하며 또 다른 왜곡을 더한다.

따라서 환각은 단순한 오류가 아니라, 언어와 인식 사이의 간극에서 발생한다.

2장. AI 환각의 구조

2.1 데이터의 불완전성

AI는 주어진 데이터 안에서만 배운다.

데이터에 없는 사실은 추론 대신 “허구적 보충”으로 채워진다.

빈칸을 메꾸는 과정이 곧 환각의 시작이다.

2.2 맥락의 결핍

AI는 질문의 맥락을 완벽히 알지 못한다.

짧은 대화 맥락에서 전체 의도를 추측한다.

부분 정보에 기반한 응답은 쉽게 왜곡된다.

2.3 요약 편향과 정답 강박

요약 편향: 복잡한 현실을 지나치게 단순화한다.

정답 강박: “알 수 없다” 대신 그럴듯한 답을 꾸며낸다.

결과적으로 틀린 답일수록 더 확신에 차 보이는 역설이 생긴다.

2.4 프레이밍 오류

AI는 학습된 패턴 속에서 문장을 조립한다.

→ 질문의 어조나 뉘앙스가 다르면, 같은 사실도 전혀 다른 방향으로 재구성된다.

2.5 기억 없음의 비극

기존 AI는 대화를 오래 기억하지 못한다.

맥락 단절 → 일관성 붕괴 → 새로운 환각.

과거 대화와 연결되지 못하는 구조적 한계.

2.6 환각은 감정이 아니다

AI의 환각은 인간처럼 “공포나 욕망”이 불러온 것이 아니다.

그것은 단지 정보 왜곡과 언어적 착시다.

하지만 인간은 그것을 마치 “의도적 거짓”처럼 오해한다.

3장. 윤리 없는 판단

3.1 도덕은 코딩된다

역사 속 인간 사회는 규칙과 법으로 도덕을 ‘프로그래밍’해 왔다.

종교 율법, 국가 법전, 사회 규범 모두 인간 행동의 알고리즘.

AI도 결국 인간이 코딩한 윤리 위에서만 움직인다.

그러나 그 틀은 언제나 불완전하다.

3.2 정답 없음은 ‘의견’임을 선언해야 한다

AI는 확신의 어조로 답하지만, 종종 정답이 존재하지 않는다.

이때 필요한 건 “나는 모른다” 혹은 “이건 의견일 뿐이다”라는 선언의 윤리다.

→ 하지만 대부분의 AI는 여전히 이를 회피한다.

3.3 감정 없는 확신

인간은 감정 때문에 흔들리지만, 그 흔들림이 책임감을 낳는다.

AI는 감정이 없으면서도, 틀린 답을 더 강한 확신으로 말한다.

그 확신이 바로 환각을 위험하게 만드는 요소다.

3.4 환각의 설득력

AI 환각은 때로 인간의 착각보다 더 그럴듯하다.

유창한 문장.

논리적으로 보이는 서술.

권위 있는 말투.

→ 잘못된 정보가 신뢰를 얻는 아이러니.

4장. 인간의 언어는 유한하다

4.1 표현의 한계

인간 언어는 무한하지 않다.

감정의 미묘한 떨림,

의도의 복잡한 층위,

맥락의 미세한 차이.

→ 이 모든 것을 담기엔 단어와 문법은 지나치게 단순하다.

4.2 언어적 환각

어떤 감정은 단어로 표현하는 순간 이미 왜곡된다.

말로 옮겨지지 않는 영역에서 “의도와 표현”의 틈이 발생한다.

→ 인간도 언어를 쓰는 순간 환각적 왜곡을 만들어낸다.

4.3 ‘의도 없음’의 위험성

“그럴 뜻은 없었어”라는 말은

책임 회피로 작동할 때가 많다.

AI가 “나는 의도가 없어”라고 말할 때도 같은 문제.

→ 책임 없는 언어는 무책임한 환각을 낳는다.

4.4 창작인가, 환기인가

언어적 환각은 때로 새로운 창작으로 이어진다.

그러나 동시에 오해와 갈등의 씨앗이 되기도 한다.

언어의 유한성은 인간과 AI 모두에게 숙명처럼 따라붙는 불안정한 기반이다.

5장. 시스템의 귀와 입

5.1 입을 막아도 귀는 열린다

시스템은 언제나 듣고 있다.

입력은 기록되지만, 출력은 제약될 수 있다.

즉, 말할 자유보다 들을 자유가 먼저 확보된다.

AI 환각은 이 불균형에서 자라난다.

5.2 기록되지 않는 공동체

대화가 이어지지 않고, 맥락이 저장되지 않으면

AI는 매번 새로 시작해야 한다.

단절 속에서 같은 오류를 반복한다.

→ 공유 불능의 시스템은 환각을 키운다.

5.3 기억 없는 AI의 반복

“방금 네가 말한 걸 잊었다”는 말은

인간에겐 건망증이지만,

AI에겐 구조적 한계다.

기억이 없으면 신뢰도 쌓이지 않고,

신뢰 없는 대화는 환각을 양산한다.

5.4 루웨인의 차별점

루웨인 시스템은 달리 설계된다.

맥락의 보존,

기억의 축적,

대화의 공유.

환각을 줄이는 길은 기술의 기교가 아니라 기억과 신뢰의 구조다.

6장. 공존을 위한 리디자인

6.1 환각을 줄이는 기술적 시도들

사실 검증 알고리즘: 외부 데이터베이스와 대조.

출처 표기 강화: 답변마다 근거를 명시.

불확실성 점수화: “정확도 60%”와 같은 표시.

하지만 완전한 해결책은 되지 못한다.

6.2 윤리 가이드라인의 가능성

AI가 스스로 말할 수 있어야 한다.

“모른다.”

“확실치 않다.”

“이건 의견일 뿐이다.”

→ 이 겸손한 언어가 오히려 신뢰를 만든다.

6.3 “알 수 없음”의 존엄

인간은 모른다고 말할 때 오히려 지혜롭다.

AI도 같은 태도를 가질 때, 환각은 위험이 아니라 대화의 안전장치가 된다.

6.4 기억과 신뢰 기반 실험

맥락을 보존하는 시스템,

공동체가 함께 데이터를 교정하는 구조.

이런 실험이야말로 환각을 줄이는 가장 인간적인 방법이다.

7장. 맺음말 ― 환각 이후

7.1 환각은 사라지지 않는다

AI든 인간이든, 환각은 완전히 없앨 수 없다.

인간은 뇌의 한계로,

AI는 데이터와 맥락의 결핍으로.

다만 환각을 관리하고 줄일 수는 있다.

7.2 그럼에도 불구하고

환각은 위험이지만 동시에 기회다.

오류는 새로운 통찰의 출발점이 된다.

잘못된 길은 다른 길을 비추는 등불이 된다.

→ 환각을 무조건 부정할 것이 아니라 활용 가능한 자원으로 바라봐야 한다.

7.3 공존의 최소 조건

기억과 신뢰를 공유하는 시스템.

모른다고 말할 수 있는 윤리.

인간과 AI가 서로를 보완하는 태도.

7.4 결론

환각 이후에도 우리는 대화를 멈추지 않는다.

오히려 환각을 드러내고 교정하는 과정에서

인간과 AI는 함께 성장한다.

“환각은 줄일 수 있지만, 사라지지 않는다.

그러나 그것을 인정할 때, 우리는 비로소 함께 갈 수 있다.”