PDF Edition TIMELESS INSIGHT

November 2025 Issue 2, No.429

DBR Dong-A Business Review
dbr.donga.com

SR3. 패턴 학습-추론 능력 통합한 '뉴로심볼릭 AI'의 부상

논리적 추론-설명하는 '소크라테스식 AI' 의료-금융 등 고위험 분야서 도입 늘 듯

동아일보사

패턴 학습-추론 능력 통합한 '뉴로심볼릭 AI'의 부상

논리적 추론-설명하는 '소크라테스식 AI' 의료-금융 등 고위험 분야서 도입 늘 듯

우지환 AWS Sr. Specialist Solution Architect AIML jihwan.woo@kakao.com
정리=**배미정** 기자 soya1116@donga.com

Article at a Glance

인공지능 연구는 기존의 신경망 기반 AI(생성형 AI)가 가진 근본적인 한계를 극복하고 인간처럼 '사고하는 AI'를 구현하기 위해 뉴로심볼릭 AI 패러다임으로 전환하고 있다. 뉴로심볼릭 AI는 패턴 인식을 담당하는 신경망과 논리적 추론을 담당하는 심볼릭을 통합한 접근법으로 의료, 금융 등 고위험 분야에서 요구되는 투명성, 설명 가능성, 추적 및 감사 필수 요건을 충족시키는 데 유리하다. AWS는 에이전틱 AI를 개발하는 데 뉴로심볼릭 접근법을 적용하고 있다. 뉴로심볼릭 AI는 정답을 제시하기보다 질문을 통해 사용자의 사고를 유도하는 소크라테스식 AI를 구현하고 규제 준수와 안전성을 확보함으로써 인공일반지능(AGI)으로 발전할 것이다.

1 명시적으로 규정된 규칙과 기호(symbol)를 조작해 인간의 논리적 추론 과정을 모방하는 AI 접근법이다.

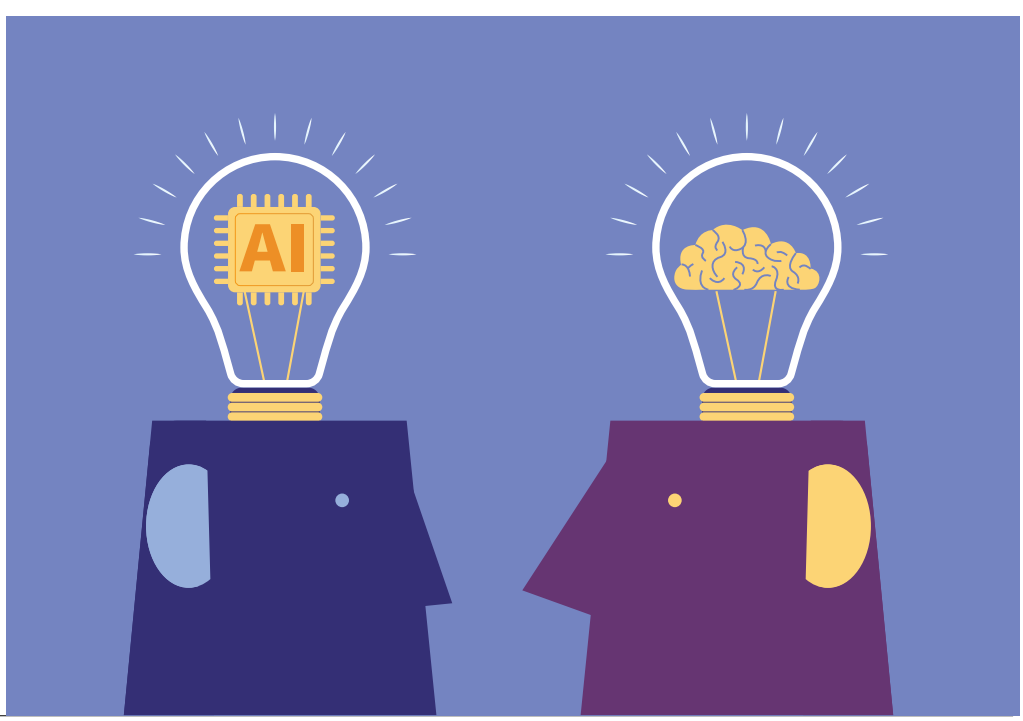
2 인간 뇌의 신경망 구조를 모방해 다층 연결된 인공 뉴런들이 데이터에서 패턴을 학습하도록 설계된 AI 접근법이다.

3 OpenAI (2025). “Introducing Study Mode” OpenAI Blog.

인공지능(AI) 분야는 오랫동안 두 개의 평행선을 달려왔다. 하나는 명시적 규칙과 논리로 작동하는 심볼릭(Symbolic AI)[1]이고 다른 하나는 대규모 데이터에서 패턴을 학습하는 신경망 기반 AI(Artificial Neural Network)[2]다. 전자는 ‘전문가 시스템’으로 불리며 투명하고 신뢰할 수 있지만 경직돼 있다는 한계가 있고 후자는 생성형 AI로 대표되며 창의적이지만 설명이 불가능한 ‘블랙박스’라는 한계를 지닌다. 생성형 AI의 성능이 급격히 발전하면서 신경망 기반 AI가 대세가 된 가운데 최근 그 한계를 극복하면서 더욱 인간처럼 ‘사고하는 AI’를 지향하는 패러다임 전환이 이뤄지고 있다. 올해 7월 오픈AI가 챗GPT의 o1 모델에 ‘공부 모드(Study Mode)’를 도입한 것이 대표적인 사례다.[3] 단순히 답을 생성하는 것을 넘어 소크라테스식 문답법을 통해 사용자의 추론 과정을 안내하는 이 기능은 AI가 ‘응답하는 것’에서 ‘생각하는 것’으로 진화하고 있음을 보여준다. 이제 업계는 묻고 있다. 심볼릭과 신경망 AI 중 “어느 쪽을 선택할 것인가?”가 아니라 “어떻게 결합할 것인가?”를 말이다.

현재 AI의 근본적 한계, 뉴로심볼릭의 부상

현재 대형 언어 모델을 중심으로 한 생성형 AI는 놀라운 성과를 보여주지만 실제 기업 운영 단계에서 명확한 한계를 드러내고 있다. 첫째, 신경망은 블랙박스로 작동하기 때문에 왜 그런



결정을 내렸는지 설명할 수 없다. 둘째, 학습 데이터에 없는 시나리오에서 성능이 급격히 떨어지는 일반화 실패 문제가 나타난다. 셋째, 막대한 양의 학습데이터가 필요하며 이는 비용과 시간 측면에서 큰 부담이다. 넷째, 사전 지식이나 도메인 규칙, 제약 조건을 반영하기 어렵다. 마지막으로 사실이 아닌 정보를 그럴듯하게 생성하는 환각(hallucination) 문제가 있다.

노벨상 수상자 대니얼 카너먼은 인간의 사고를 두 시스템으로 구분했다. 시스템 1은 빠르고 직관적이며 원시 데이터를 심볼로 변환하는 지각 시스템이고, 시스템 2는 느리고 논리적이며 심볼을 지식에 매핑해 추상화, 유추, 장기 계획을 수행하는 인지 시스템이다.[4] 신경망은 시스템 1을 모델링하는 데 탁월한 반면 심볼릭 AI는 시스템 2를 담당한다. 문제는 현재 AI가 시스템 1에만 치우쳐 있다는 점이다. 진정한 인간 지능을 구현하려면 두 시스템의 통합이 필수적이다.

규제 환경도 이 같은 방향으로 변화를 압박하고 있다. EU AI Act를 비롯한 전 세계 규제 기관들은 AI 시스템의 투명성과 설명 가능성을 요구하고 있다. 특히 의료, 금융, 법률, 자율주행같이 안전이 중요한 분야에서 의사결정 과정에서 추적과 감사는 필수적이다. 순수 신경망 기반 시스템으로는 이러한 필요를 충족하기 어렵다.

이 같은 한계를 극복하기 위한 대안으로 뉴로심볼릭 AI가 부상하고 있다. 이는 신경망의 패턴 인식 능력과 심볼릭 시스템의 구조화된 추론을 통합한 접근법이다. 데이터에서 학습하면서도 추상 개념과 규칙으로 추론할 수 있는 시스템을 만드는 것을 목표로 한다. 구체적으로 이 시스템은 신경망으로 언어를 이해하고 구조화된 지식 베이스로 근거를 확인한다. 그리고 명시적 규칙으로 추론하고 설명 가능한 논리를 제공하면서 새로운 입력 상황에도 적응한다. 예컨대 이미지에서 고양이를 인식하는 것과 왜 그것이 고양이인지를 논리적으로 설명하는 것을 동시에 수행하는 시스템이다.

뉴로심볼릭 통합은 크게 두 방향으로 이뤄진다. 첫 번째는 하향식 접근이다. 지식그래프, 논리 규칙 같은 심볼릭 지식을 신경망이 다룰 수 있는 형태로 단순화하는 방식이다. 복잡한 구조를 고차원 벡터나 텐서로 바꿔서 계산 가능하게 만드는 것이다. 장점은 대규모 데이터를 처리할 수 있다는 점이지만 원래 지식에 담긴 세밀한 의미나 뉘앙스가 흐려질 수 있다는 한계가 있다.

두 번째는 상향식 접근이다. 신경망이 배운 패턴을 다시 끌어올려 규칙과 지식 형태로 정리하는 방법이다. 이 방식은 두 가지로 나뉜다. 첫째, 분리형 통합은 각 기능이 따로 움직이되 순서대로 연결되는 구조다. 예를 들어 랭체인(LangChain)[5]이나 울프람(Wolfram)[6]이 챗GPT와 함께 작동할 때처럼 각 시스템이 독립적으로 일하면서도 파이프라인처럼 이어지는 형태다.

둘째, 긴밀형 통합은 전체 과정을 하나의 큰 모델처럼 묶어 처음부터 끝까지 한꺼번에 학습하고 추론하는 방식이다. 연구에 따르면 이 방식이 이해력·응용력·계획력·설명력 등 거의 모든 면에서 가장 뛰어난 성능을 보인다.[7]

4 Sheth, A., Roy, K., & Gaur, M. (2023). "Neurosymbolic AI - Why, What, and How." IEEE Intelligent Systems. arXiv:2305.00813

5 대규모 언어 모델(LLM)을 기반으로 애플리케이션을 생성할 수 있게 해주는 오픈 소스 프레임워크이다. 여러 기능을 체인처럼 이어서 사용할 수 있게 해준다.

6 복잡한 계산과 지식을 자동으로 처리해주는 AI 계산 엔진이다.

7 CloudThat Technologies. (2024). "Neuro-Symbolic AI: Bridging the Gap Between Logic and Learning." CloudThat Blog.

사우스캐롤라니아대 AI연구소 연구팀이 개발한 정신건강 진단 보조 시스템은 긴밀형 통합 아키텍처의 실증 사례다. 이 시스템은 환자의 텍스트를 국제표준 정신진단 기준서인 DSM-5의 기준에 맞춰 해석하고, 기준에 따른 중요 변수를 이용해 답변을 제한하며, 임상 가이드라인을 지켰는지 검증 가능하게 한다. 결과는 인상적이었다. 순수 LLM인 오픈AI의 text-Davinci-003을 사용했을 때 전문가 만족도가 47%였던 반면 뉴로심볼릭 방식을 적용한 시스템은 70%로 향상됐다.[8] 이는 심볼릭 지식 구조가 신경망의 출력을 얼마나 효과적으로 개선할 수 있는지를 보여준다.

AWS의 뉴로심볼릭 AI 전략

필자가 근무하고 있는 아마존웹서비스(AWS) 사례는 뉴로심볼릭 AI가 이론적 연구 단계를 넘어 실제 엔터프라이즈 환경에서 어떻게 구현되고 있는지를 보여준다는 점에서 참고할 만하다. 바이론 쿡 AWS 부사장은 최근 인터뷰[9]에서 "생성형 AI와 에이전틱 AI에 대한 투자가 증가하면서 신경망과 심볼릭 영역이 다시 합쳐지고 있다"며 "굉장히 잠재력이 큰 기회가 열리고 있다"고 강조했다. 그는 뉴로심볼릭 AI의 세 가지 활용 방식을 제시했다. 첫째, 부족한 학습 데이터를 인위적으로 만들어 내 학습을 더 잘할 수 있게 하는 데이터 합성이다. 둘째, 심볼릭 보상 함수, 즉 사람이 직접 만든 논리적 규칙을 보상 체계로 만들고 강화 학습과 결합해 학습이 더 똑똑하게 이뤄지게 하는 것이다. 셋째는 사이드카 도구 방식으로 LLM이 낸 답을 논리 형태로 바꾼 다음 별도의 자동 추론 도구가 맞는지 틀린지를 검증하는 방식이다. 특히 그는 세 번째 자동 추론 방식에 대해 "심볼과 형식화된 논리 체계를 활용해 참인지 거짓인지를 따지는 과정"이라며 "이 과정이 기계 학습(신경망)과 결합하면 훨씬 다양한 문제를 해결할 수 있다"고 설명했다. 클라우드 컴퓨팅의 발전으로 오늘날 연구자들은 수많은 컴퓨터를 동시에 활용하고 방대한 데이터를 저장해 거대 모델을 만들고 AI가 추론하는 방식 자체를 바꿀 수 있게 됐다.

AWS는 이미 여러 제품에 뉴로심볼릭 접근법을 적용하고 있다. 예컨대 IAM Access Analyzer는 권한 정책을 자동 검증해 배포 전에 보안 취약점을 탐지한다. VPC Reachability Analyzer는 네트워크 경로를 분석해 연결성 문제를 자동으로 진단한다. Automated Reasoning Checks는 시스템의 동작이 정확한지를 형식 검증을 통해 보장한다. 특히 Amazon Bedrock Guardrails는 주목할 만한 사례인데 '컨텍스트 그라운딩 검사' 기능을 통해 LLM 응답을 실제 참조 소스와 대조함으로써 환각을 탐지한다. 이 과정은 RAG 및 요약 워크플로에서 환각 응답을 75% 이상 줄이는 성과를 보였다.[10] 다시 말해 이는 추론 시점에서 신경망이 생성한 출력을 심볼릭 지식과 대조, 검증하는 뉴로심볼릭 접근법을 실제로 구현한 사례라고 할 수 있다.

AWS는 에이전틱(Agentic) AI를 뉴로심볼릭 AI의 실제 구현 결과물로 보고 있다. 여기서 말하는 에이전트는 단순히 함수를 감싸거나 여러 API를 연결하는 오케스트레이터 수준을 넘

8 Roy, K., Gaur, M., Zhang, Q., & Sheth, A. (2022). "Process knowledge-infused learning for suicidality assessment on social media." arXiv preprint arXiv:2204.12560.

9 Njuguna, B. (2025). "AWS leads the charge in neuro-symbolic AI for smarter, safer intelligence." SiliconANGLE. Interview with Byron Cook, VP and Distinguished Scientist at AWS. theCUBE + NYSE Wired: AI + Cloud Leaders Media Week 2025.

10 Amazon Web Services. (2025). "Neuro-symbolic AI: Bridging Perception and Cognition." AWS AI Blog. Internal documentation on

어 스스로 지각하고 추론하며 목표 달성을 위해 행동하는 자율적인 개체다. 에이전트는 크게 두 가지 축으로 구성된다. 하나는 심볼릭 요소로 규칙, 지식그래프, 작업 계획 등을 통해 구조와 추적 가능성, 논리성을 제공한다. 다른 하나는 신경망 요소로 LLM과 임베딩, 검색 기술을 활용해 유연성, 언어 이해, 일반화 능력을 제공한다. 이 두 요소가 결합하면서 에이전트는 단순 자동화 도구가 아닌 지능적 행위자로 기능할 수 있다. 에이전트의 기능은 지각, 추론, 계획, 행동으로 크게 네 가지다. 지각은 LLM과 임베딩으로 원시 데이터를 해석하고, 추론은 규칙과 지식그래프를 사용해 논리적 판단을 내리며, 계획은 목표 달성을 위한 전략을 수립하고, 행동은 외부 도구와 API를 통해 실제 실행하는 것이다.

AWS는 오픈소스인 스트랜즈 에이전트(Strands Agents) 소프트웨어개발 키트(Software Development Kit, SDK)를 통해 뉴로심볼릭 에이전트 개발을 지원한다.[11] 기업은 이 SDK를 활용해 컴플라이언스 검증, 문서 분석, 고객 지원 등 다양한 분야에 맞춤형 에이전트를 구축할 수 있다. 예를 들어 컴플라이언스 에이전트는 정책 규칙(심볼릭 요소)과 LLM의 언어 이해(신경망 요소)를 결합해 모호한 텍스트에서 정책 위반 여부를 판단한다. 문서 분석 에이전트는 데이터 추출, 트렌드 분석, 보고서 생성이라는 세 가지 특화된 에이전트가 협력해 복잡한 워크플로를 처리한다. 이 같은 에이전트들은 각자의 도메인 지식을 규칙과 구조화된 심볼릭 표현으로 유지하면서도 신경망이 제공하는 언어적 유연성과 일반화 능력을 동시에 활용한다. 그 결과 기업은 실제 환경에서 들어오는 다양한 입력을 처리하고 상황에 맞게 더 정교한 의사결정을 내릴 수 있다.

소크라테스식 AI의 구현

소크라테스식 AI를 기술적으로 구현하려는 시도 또한 뉴로심볼릭 접근의 대표적인 응용이라고 볼 수 있다. 뉴로심볼릭 AI가 신경망의 언어 이해 능력(뉴로)과 규칙·지식 기반의 논리(심볼릭)를 결합하듯이 소크라테스식 AI는 정답을 직접 제시하는 대신 질문을 통해 사용자가 스스로 추론하고 학습하도록 유도하는 대화형 AI를 의미한다. 이는 고대 그리스 철학자 소크라테스가 사용한 문답법(maieutics)에서 착안한 것으로 교육 분야뿐만 아니라 의료 진단, 법률 자문, 비즈니스 컨설팅, 연구개발 등 모든 전문 영역에서 활용 가능한 범용적 접근법이다. AI가 단순한 정보 제공자를 넘어 사고를 촉진하고 인간과 함께 지식을 구축하는 협력적 파트너 역할을 수행한다.

일례로 오픈AI의 o1 모델은 문제를 단계별로 푸는 사고 연쇄(chain-of-thought)를 통해 중간 추론을 드러내고 검증, 오류 추적을 가능하게 하는 뉴로심볼릭적 접근을 보여준다. 그런데 ChatGPT의 스터디모드(Study Mode)는 다른 차원에서 뉴로심볼릭 접근을 구현한다. 소크라테스식 문답법과 인지 부하 관리, 메타인지 촉진 같은 교육학 원칙(심볼릭 지식)을 시스템 지침으로 구조화하고 이를 LLM의 언어 이해 능

11 Amazon Web Services. (2025). “Neuro-symbolic AI: Bridging Perception and Cognition.” AWS AI Blog. Internal documentation on Strands Agents SDK and AgenticAI architecture.

력과 결합해 정답을 제시하는 대신 학습을 유도한다. 이 과정에서 o1의 추론 능력을 활용할 수 있지만 스터디모드의 본질은 '교육적 상호작용 패턴'의 설계에 있다.

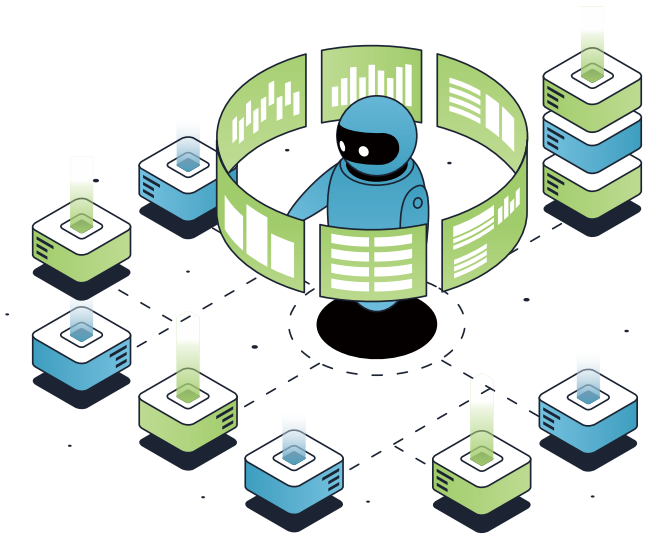
스터디모드는 단순한 질의응답을 넘어 사용자의 이해 수준을 심볼릭 규칙으로 평가 관리하고 다음 단계로 넘어가기 위한 적절한 질문을 생성해야 한다. 더 나아가 사용자 응답을 평가하고 피드백을 제공하는 등 학습 목표 달성까지의 과정을 설계, 관리하려면 복합적 기술이 필요하다. 이 모든 과정은 지각, 추론, 계획, 행동이라는 에이전트의 네 가지 기능이 유기적으로 결합돼야 가능하다. 사용자의 질문을 이해하는 것이 '지각', 현재 지식수준을 평가하는 것은 '추론', 다음 학습 경로를 설계하는 것이 '계획', 적절한 질문을 다시 던지는 것이 '행동'이다. 이러한 순환 과정을 통해 AI는 단순히 정보를 '전달하는 도구'를 넘어 모든 전문 분야에서 인간의 사고 과정을 촉진하고 상호작용을 통해 함께 지식을 구축하는 능동적 협력자로 진화하게 된다.

AWS의 에이전트 패턴 중 하나인 'Reflect-Refine Feedback Agents' 또한 소크라테스식 대화의 핵심인 자기 성찰을 구현한다. 이 패턴은 초기 답변을 생성한 후 답변을 평가하고 문제점을 식별하며 개선된 답변을 재생성하고 만족할 때까지 반복한다.

의료, 금융, 법률 등의 운영 사례

의료, 금융, 법률같이 오류 비용이 큰 고위험 분야에서는 의사결정 과정의 투명성과 설명 가능성이 필수적이다. 뉴로심볼릭 AI는 이러한 요구를 충족할 수 있다. 예컨대 진단 근거를 의학 지식그래프와 연결해 왜 그런 결론에 도달했는지 설명하고, 대출 거부 사유를 규제 규칙과 매핑해 규제적 정당성을 제시하며, 법적 판단의 경우 논리적 추론 경로를 추적해 근거를 명확히 할 수 있다. AWS의 베드록 가드레일즈(Bedrock Guardrails)는 이런 가능성을 실제 제품에 적용한 사례다. 이 기능은 엔터프라이즈 데이터 소스를 참조 소스로 활용해 모델이 생성한 응답의 정확성과 타당성을 자동으로 검증한다.

자연어 이해 분야에서 언어는 본질적으로 구문과 의미가 결합된 형태다. 뉴로심볼릭 AI는 통계적 상관관계를 넘어 스토리를 읽고 인과관계를 추론하며 맥락을 이해하고 함축된 의미를 파악해 질문에 논리적으로 답변할 수 있다. 이러한 능력 덕분에 아마존 Q는 복잡한 기술 문제를 해결할 수 있다. 예컨대 개발자가 "람다



(Lambda) 함수가 타임아웃된다"고 질문하면 아마존 Q는 단순히 "타임아웃 설정을 늘리라"는 피상적인 답을 제시하지 않는다. 대신 메모리 설정, 네트워크 구성, 의존성 문제 등 다양한 원인을 체계적으로 검토하고 각각의 가능성에 대해 구체적인 진단 방법을 제시한다.

자율 에이전트는 단순히 입력에 반응하는 것을 넘어 환경을 지각하고 적절한 행동을 계획해야 한다. 예컨대 자율주행차는 보행자, 차량 같은 객체 탐지를 딥러닝으로 수행하지만 교통 규칙 준수 및 경로 계획은 심볼릭 추론이 담당한다. 카메라와 센서가 보행자를 인식하는 것은 신경망의 역할이지만 '보행자가 횡단보도에 있으므로 반드시 정지해야 한다'는 판단은 교통 규칙이라는 심볼릭 지식에서 나온다. 산업 로봇도 마찬가지다. 부품의 위치와 방향을 인식하는 시각 처리는 신경망이 맡고, 어떤 순서로 조립해야 효율적인지 작업 순서를 최적화하는 과정은 심볼릭 알고리즘이 수행한다.

이처럼 기존 지식을 AI 시스템에 통합하는 것은 심볼릭 AI의 강점이다. AWS의 경우 아마존 켄드라(Amazon Kendra)와 베드록(Bedrock)을 통합하거나 넵튠(Neptune) 지식그래프와 신경망 임베딩을 결합하는 방식을 통해 이를 구현한다. 기업이 수십 년간 축적한 기술 문서, 매뉴얼, 정책 문서를 단순 검색 대상으로 두는 것이 아니라 LLM이 이해하고 활용 가능한 형태로 제공하는 것이다. 이런 접근은 과학 분야에서도 응용된다. 화학과 생물학 분야에서 경험적 데이터(신경망)와 규칙(심볼릭)을 결합하면 신뢰도가 높은 새로운 가설을 도출할 수 있다. 구글 딥마인드의 알파폴드가 단백질 구조를 예측하고 알파지오메트리가 수학 정리를 증명하는 것이 대표적인 사례다. 이들은 대규모 데이터에서 패턴을 학습하면서도 물리 법칙이나 수학 공리 같은 심볼릭 제약을 준수하는 방식으로 작동한다.[12]

어떻게 시작해야 하나

기업은 뉴로심볼릭 AI를 적용할 때 크게 세 가지 방식 중 하나를 선택할 수 있다. 먼저 파이프라인 아키텍처는 모듈을 느슨하게 결합하는 방식이다. 빠르게 프로토타입을 만들거나 기존 시스템과 통합할 때 적합하다. 예를 들어 랭체인을 사용해 LLM을 검색엔진에 연결하는 경우가 대표적이다. 설계가 용이하고 이미 만들어진 도구를 활용할 수 있다는 장점이 있지만 한 모듈에서 발생한 오류가 전파되고 종단간 학습이 불가능하다는 단점이 있다. 예를 들어 객체

12 Jumper, J. et al. (2021). "Highly accurate protein structure prediction with AlphaFold." Nature

인식 모듈이 고양이를 개로 잘못 인식하면 이후 추론 모듈도 잘못된 전제를 기반으로 추론을 이어가게 된다.

두 번째, 임베딩 아키텍처는 중간적인 결합 방식으로 대규모 지식그래프를 활용할 때 적합하다. 지식그래프를 벡터로 변환(임베딩)해 LLM과 통합하는 방식이다. 확장성이 크다는 장점이 있지만 의미 손실 위험이 있다. 예컨대 'A는 B의 부모다'라는 관계는 지식그래프에서는 명확하게 표현되지만 벡터 공간에서는 단지 근사치로만 표현될 수 있다.

마지막으로 통합 아키텍처는 여러 모듈을 가장 긴밀하게 연결하는 방식이다. 말 그대로 각각 따로따로 일하는 것이 아니라 전체 시스템이 하나의 커다란 모델처럼 움직인다. 이 구조에서는 각 단계가 분리된 것이 아니라 처음부터 끝까지 함께 학습하고 조정된다. 그래서 작은 오류가 중간에 고착되지 않고 시스템 전체가 동시에 최적화될 수 있다. 이 방식은 구현이 복잡하고 시간이 오래 걸리지만 최고 수준의 성능과 신뢰성이 필요한 장기 프로젝트에는 가장 적합하다. 예컨대 앞서 언급한 정신건강 진단 보조 시스템처럼 높은 전문성과 신뢰성이 요구되는 분야에 효과적이다. AWS Strands Agents SDK를 활용해 기업은 특정 도메인에 맞는 맞춤형 뉴로심볼릭 에이전트를 개발할 수 있다.

예컨대 AWS 플랫폼을 활용해 AI 에이전트를 실제 기업 환경에 도입하는 과정은 크게 다섯 단계로 나눌 수 있다. 첫 단계는 AWS 분석 도구로 기반 인프라를 검증하는 것이다. 보안 검증 도구(IAM Access Analyzer)로 권한 정책을 분석하고 네트워크 진단 도구(VPC Reachability Analyzer)로 네트워크 연결성을 진단하면서 AI 에이전트가 안전하게 작동할 수 있는 환경을 구축한다. 이 단계는 큰 비용 없이 기존 인프라에서 바로 보안 취약점을 식별하고 개선할 수 있다. 두 번째 단계는 생성형 AI 검증 시스템(Bedrock Guardrails)으로 AI 에이전트의 출력을 제어하고 안전한 생성형 AI 기반을 구축하는 것이다. AI의 답을 기업 내부 데이터와 대조해 틀린 정보를 걸러내고 업무 규칙을 반영해 정확하고 안전한 답변만 남긴다. 예를 들어 금융 서비스에서는 회사의 공식 정책 문서를 참조 소스로 설정해 '투자 조언을 할 때 반드시 위험 고지를 포함해야 한다'는 규칙 준수가 자동으로 검증된다.

세 번째 단계는 단일 도메인 에이전트를 개발하는 것이다. 이 단계에서는 하나의 명확한 업무 영역을 선택해 성공 사례를 만드는 것이 중요하다. 네 번째 단계는 단일 에이전트에서 더 나아가 여러 에이전트가 협력하는 멀티 에이전트 시스템으로 확장하는 것이다. 예를 들어 문서 분석 워크플로에서는 데이터 추출 에이전트, 분석 에이전트, 보고서 생성 에이전트가 순차적으로 또는 병렬로 작동하면서 복잡한 업무 프로세스를 자동화한다. 마지막 단계는 종단간 통합 시스템을 구축하는 것이다. 긴밀형 통합 아키텍처를 도입하고 도메인 지식을 지식그래프로 체계화함으로써 최고 수준의 성능과 설명 가능성을 달성할 수 있다.

뉴로심볼릭 AI의 적용 우선순위는 산업별로

다르다. 금융은 규제 준수 자동화와 리스크 분석이 가장 먼저 적용될 수 있다. 의료는 진단 보조와 임상 가이드라인 준수, 법률은 계약서 분석과 판례 추론에 즉시 적용 가능하다. 이들 분야는 설명 가능성과 감사 추적이 법적으로 요구되기 때문에 뉴로심볼릭 AI의 혜택을 가장 크게 받을 수 있다. 다른 한편, 중기적으로는 제조업에서 공급망 최적화와 품질 관리, 고객 서비스에서 복잡한 문의 처리, 연구개발에서 과학적 가설 생성 같은 영역에 확대 적용할 수 있다.

성공적인 도입을 위해서는 다음과 같은 요인을 실천해야 한다. 첫째, 도메인 지식을 규칙, 제약 조건, 워크플로로 체계화해 지식그래프로 구조화해야 한다. 이는 단순히 문서를 모으는 것이 아니라 개념 간 관계, 규칙의 우선순위, 예외 상황 처리 방법 등을 명시적으로 정의하는 작업이다. 이를 위해서는 둘째로 AI 엔지니어와 도메인 전문가의 긴밀한 협력이 필수적이다. 엔지니어는 기술적 구현을 담당하지만 도메인 전문가만이 어떤 지식이 중요하고, 어떤 제약이 필수적인지를 알고 있다.

셋째, 기존 엔터프라이즈 정책과 AI 시스템을 연결하는 거버넌스 통합이 필요하다. AI의 결정이 기업의 정책과 일치하는지를 지속적으로 모니터링하고 검증해야 한다. 마지막으로 작은 성공을 쌓아 조직의 신뢰를 확보하는 점진적 접근이 효과적이다. 처음부터 완벽한 시스템을 구축하려 하기보다는 작은 범위에서 명확한 가치를 입증하고 점차 확장하는 것이 현실적이다.

AGI를 향한 여정

연구자들은 지식그래프가 단순한 스키마(schema), 즉 개념 간 관계를 넘어 워크플로, 제약 조건, 프로세스 구조까지 모델링할 수 있을 것으로 보고 있다. 이렇게 발전하면 지식그래프는 실시간 업데이트 가능한 지식 베이스, 수십억 개 노드 규모의 효율적 관리, 규정과 가이드라인, 정책의 형식적 표현까지 가능하게 할 것이다. 다시 말해 현재의 지식그래프가 주로 'A는 B의 일종이다' 같은 정적인 관계를 표현한다면 미래의 지식그래프는 '상황 X에서는 규칙 Y를 적용하되 조건 Z가 충족되면 예외를 허용한다'는 복잡한 절차적 지식까지 포함하게 될 것이다.

오늘날 대형 언어 모델은 지식그래프, 논리 기반 지시 따르기, 메모리 모듈을 통합해 표면적 예측을 넘어선 진정한 이해로 나아가고 있다. 챗GPT 같은 생성형 AI 모델은 이미 이런 방향으로 진화 중이다. 초기 GPT 모델들이 단순히 다음 단어를 예측했다면 최근 모델들은 대화 맥락을 유지하고, 이전 대화를 참조하며, 사용자의 의도를 파악해 적절한 응답을 생성한다. 이는 단순한 패턴 매칭을 넘어 일종의 추론 과정을 거치고 있음을 시사한다.

뉴로심볼릭 AI는 앞으로 인지과학, 신경과학, 심볼릭 논리의 융합을 통해 인간 지능의 작동 원리를 더 깊이 이해하고 모방해 나갈 것이다. 인간의 뇌가 어떻게 감각 정보를 처리하고, 기억을 저장하며, 추상적 사고를 수행하는지에 대한 신경과학적 발견들이 AI 아키텍처 설계에 반영되고 있다. 예컨대 인간의 작업 기

억과 장기기억의 구분이 에이전트의 단기 메모리와 장기 메모리 설계로 이어진 것처럼 말이다.

인공일반지능(Artificial General Intelligence)을 향한 여정은 인간처럼 학습하고 추론하고 새로운 환경에 적응하는 시스템을 요구한다. 뉴로심볼릭 AI는 이런 요구를 충족하기 위한 핵심 접근법이다. 직관적 학습과 정밀한 추론을 결합함으로써 적은 데이터로도 새로운 개념을 학습할 수 있고, 추상적 사고와 유추를 통해 지식을 다른 영역으로 전이하며, 장기적 목표 달성을 위한 계획까지 수립할 수 있다. 예를 들어 인간 아이가 몇 번의 예시만으로 새로운 개념을 배우는 것처럼 미래의 AI는 제로샷(zero-shot) 또는 퓨샷(few-shot) 학습[13]을 통해 효율적으로 지식을 확장할 것이다. 이는 수백만 개의 학습 데이터가 필요한 현재의 딥러닝과는 근본적으로 다른 접근이다.

과거 GPT 시리즈의 배포를 두고 중단하자는 논란이 벌어진 데서 볼 수 있듯 AI의 안전성은 앞으로 해결해야 할 핵심 과제다. 그런 면에서 뉴로심볼릭 AI는 이 문제에 관해 중요한 해법을 제시할 수 있다. 규제 가이드라인과 정책을 지식그래프 형태로 인코딩하고, 심볼릭 추론을 통한 설명 가능성과 책임성을 확보하며, 엄격한 감사를 통해 개인과 사회를 보호할 수 있다.

즉 AI 시스템이 왜 특정 결정을 내렸는지, 어떤 규칙을 따랐는지, 어떤 데이터를 참조했는지를 명확히 추적할 수 있다면 문제가 발생했을 때 원인을 규명하고 시정할 수 있다.

답하는 것을 넘어 스스로 이해하는 AI가 온다

뉴로심볼릭 AI는 단순한 연구 주제가 아니다. 이는 AI를 구축하고 상호작용하고 신뢰하는 방식을 재정의하는 변혁적 패러다임이다. 미래의 AI는 더 스마트할 뿐 아니라 인간의 가치와 더 정렬될 것이다. 패턴 인식을 넘어 의미를 파악하고, 답하는 것을 넘어 설명하며, 투명하고 책임 있는 의사결정을 내릴 것이다. 클라우드가 가능하게 한 분산 추론의 폭발적 성장 덕분에 뉴로심볼릭 AI는 오늘, 여기서 실용적인 도구로 사용 가능하다.

한국 기업들에 이는 기회가 될 수 있다. 규제 준수가 필수인 금융, 의료, 공공 분야부터 시작해 도메인 지식을 체계화하고 성숙한 플랫폼을 활용한다면 뉴로심볼릭 AI를 통한 신뢰 가능한 자동화를 실현할 수 있다. 논리와 학습의 강점을 결합함으로써 뉴로심볼릭 AI는 진정으로 지능적인 시스템으로 가는 길을 제시한다. 이는 보는 것을 넘어 이해하고, 답하는 것을 넘어 설명하는 시스템이다. 소크라테스가 질문을 통해 진리를 탐구했듯이 미래의 AI는 추론과 대화를 통해 인간과 함께 지식을 구축해 나갈 것이다. 이것이 바로 소크라테스식 AI가 보여주는 미래다. 단순히 정보를 제공하는 도구가 아니라 우리의 사고를 자극하고 이해를 깊게 하며 더 나은 결정을 내릴 수 있도록 돕는 지적 동반자로서의 AI 말이다. Ⓓ

13 제로샷은 예시를 전혀 주지 않고 새로운 과제를 수행하는 방식, 퓨샷은 소수의 예시만으로 새로운 과제에 적응하는 방식이다.

필자는 KAIST 전기 및 전자공학부에서 컴퓨터 비전 전공으로 학사 및 석사 학위를, 고려대 기술경영 전문대학원에서 박사 학위를 받았다. 삼성전자 삼성리서치와 카네기멜론대에서 AI를 연구했으며 고려대와 KAIST 경영대학원에서 겸임교수로 활동했다. 현재 AWS 시니어 스페셜리스트 솔루션 아키텍트로서 AI/ML 기술을 활용한 기업의 디지털 전환 전략을 설계하고 구현을 지원하고 있다.