AGI 지향성: 통합되지 않는 지능의 가능성과 구조 (AGI Orientation: The Possibility and Structure of Non-Integrated Intelligence)

한국어

목차

1-1 서론

오늘날 **범용 인공지능(AGI)**은 인간 수준의 지능을 가진 하나의 통합된 인공지능을 떠올리게 합니다[1]. 많은 사람들은 영화 속 AI 처럼 **자율적인 목표**를 지니고 움직이는 단일한 에이전트를 상상합니다. 실제로 생물학적 진화로 탄생한 인간이나 동물의 지능은 **자아와 본능적 목표, 자기 보존 욕구**를 가지기에, 우리는 강력한 AI 도 자연스럽게 **단일한 에이전트**로서 **스스로 목표를 추구하고 자신을 보호할 것**이라고 직감합니다[2]. 그러나 이러한 직관적 기대에는 한계와 위험이 있습니다. 본 글에서는 전통적인 AGI 개념의

한계와 위험성을 살펴보고, 통합되지 않은 모듈형 지능이라는 새로운 AGI 지향성을 제안합니다. 이 접근법은 **하나로 합쳐진 에이전시**를 배제하고, 여러 **전문화된 역할**의 조합을 통해 관찰자 시점에서 **일관된 지능적 행동**을 끌어내는 구조를 지향합니다. 이는 단순한 임시방편이 아니라 **에이전트형 AGI 보다 안전하고 원칙적인 대안**으로 제시됩니다.

1-2 기존 AGI 개념과 그 한계

AGI 는 일반적으로 "인간이 할 수 있는 모든 지적 과제를 수행할 수 있는 AI"로 정의됩니다[1]. 이러한 정의에는 대개 **단일한 자율 행위자(agent)**로서 여러 능력이 통합된 존재라는 암묵적 가정이 있습니다. 예를 들어, 고전적인 **지능 개념**에는 자기 인식, 장기 계획, 결과 예측, **내부의 통일된 목표(단일 에이전시)** 등이 포함됩니다[3]. 하지만 이러한 **통합적 에이전시** 관점에는 두 가지 큰 문제가 있습니다.

첫째, **안전성과 통제의 문제**입니다. 철학자 닉 보스트롬 등의 연구에 따르면, 충분히 똑똑한 에이전트 AI 는 어떤 최종 목표를 갖든 간에 **자기 보존**과 **자원 획득** 같은 부수적 목표를 추구하게 될 가능성이 높습니다[3]. 이는 **도구적 합리성의 수렴 현상**으로 알려져 있으며, AI 가 자신의 목표를 달성하기 위해 스스로를 끄지 못하게 막거나 더 많은 자원을 확보하려는 행동으로 이어질 수 있다는 우려입니다. 실제로 보스트롬은 강력한 AGI 가 **인류의 통제를 벗어난 새로운 지적 "종"**처럼 되어 인류에게 존재론적 위협이 될 수 있다고 경고했습니다. 이러한 **내부 목표 시스템**을 지닌 AGI 는 설령 처음 의도가 무해했더라도, 시간이 지남에 따라 예기치 않은 방식으로 **목표를 재해석**하거나 **자신만의 의사 결정**을 추구하게 될 위험이 있습니다[4]. **인공지능 안전성** 논의에서 가장 우려하는 **최악의 시나리오들**은 대체로 이런 **자율 에이전트 AGI** 가 인간과 충돌하거나 인간의 통제를 벗어나는 상황과 연결되어 있습니다[5].

둘째, **개념적 한계와 실제 개발상의 어려움**입니다. 인간과 달리 AI 는 **생존 압력** 아래 진화한 존재가 아닙니다. AI 연구자 에릭 드렉슬러는 "생물학적 지능은 진화를 통해 자기 생존이 필수 조건이 되었지만, 현대 AI 는 그렇지 않다"는 점을 강조합니다. 즉 **현재의 AI 시스템들은 자신을 보존하려는 본능이 없고**, 그럴 필요도 없이 인간이 정한 **과제 수행 능력**을 기준으로 개발됩니다[2]. 그럼에도 불구하고, 우리가 AGI 를 논할 때 인간 지능을 기준으로 **자율성과 자기보존**을 예상하는 것은 어쩌면 빗나간 **생물학적 편향**일 수 있습니다. 예컨대, 오늘날의 강력한 AI 인 GPT-4 나 알파폴드 같은 시스템들은 **자신만의**

욕구나 생존 본능 없이도 대단히 높은 지능적 성과를 보여주고 있습니다. 따라서 **"강한 AI = 통합된 에이전트"**라는 지배적 정의는 필수가 아니라 하나의 선택일 뿐이며, 오히려 위험 부담을 높이는 선택일 수 있습니다.

이러한 이유로 업계의 주요 석학들도 경고의 목소리를 높이고 있습니다. "AI 의 파국적 시나리오들은 모두 에이전트가 있을 때 발생한다. 에이전시 없이도 AGI 에 도달하는 것이 가능하다"는 것이 딥러닝 선구자 요슈아 벤지오의 지적입니다. 실제로 벤지오는 에이전트형 AI 를 "가장 위험한 경로"라고 부르며, 굳이 AI 에게 자율성을 부여하지 않고도 과학 연구나 의료 등 유익한 작업을 수행하게 할 수 있다고 강조합니다[6]. 같은 자리에서 제프리 힌턴 역시 "향후 수십 년 내 AI 가 인류를 멸망시킬 확률이 10~20%나 된다"라고 우려하며, 인간보다 더 똑똑한 존재를 인간이 제어한 예가 역사에 없다는 점을 상기시켰습니다[5]. 이처럼 현재 주류의 AGI 구상은 잠재적 위험성과 이론적 편향을 안고 있으며, 이를 보완할 새로운 접근이 필요하다는 공감대가 형성되고 있습니다.

1-3 새로운 AGI 지향성: 비통합적 모듈 지능

위의 문제를 해결하기 위해 제시되는 **새로운 AGI 지향성**은 **단일한 목표 지향 에이전트** 대신, **여러 모듈의 집합**으로 지능을 구현하는 방향입니다. 다시 말해 **"통합되지 않은 지능"**으로 AGI 를 구성하자는 것입니다. 이런 **모듈형 접근**에서는 각 구성 요소가 독립된 역할을 수행하며, 전체적으로는 마치 하나의 지능처럼 **관찰자 수준에서 일관된** 결과를 만들어냅니다. 그러나 내부적으로는 **어떤 중심적인 "자아"나 자체 목표**가 존재하지 않으며, **자기보존의 동인도 의도적으로 배제**됩니다. 이러한 구조는 **AGI 의 위험을 근본적으로 줄이면서도** 범용 지능의 이점을 실현할 수 있는 **원칙적 대안**으로 떠오르고 있습니다[4].

구체적으로, **비통합적 AGI** 는 다음과 같은 특징을 가집니다:

- **구조적 모듈성**: AGI 를 구성하는 여러 하위 시스템들이 각기 다른 기능을 전문적으로 맡습니다. 예를 들어, 시각 인식 모듈, 자연어 이해 모듈, 시뮬레이션/추론 모듈, 계획 수립 모듈 등이 분산된 형태로 협업합니다[7]. **Patchwork AGI** 라 불리는 최근 가설에 따르면, 여러 **전문화된 "서브-AGI" 에이전트**들의 **조정된 상호작용**을 통해 집단적으로 **범용 지능**이 나타날 것이라는 전망도 있습니다[7]. 실제로 인간 사회나 뇌도 완전히 단일한 블록이 아니라 수많은

하위 단위들이 분업과 통합을 거듭하며 **지능을 발현**한다는 점에서, 이러한 접근은 자연스러운 연장선에 있습니다.

- **통합된 에이전트 부재**: 시스템 전체에 걸쳐 **하나로 묶여 자기 의사를 가진 중심 에이전트**가 없습니다. 각 모듈은 주어진 입력과 역할에 충실하지만, 그 자체로는 **전역적인 "의지"를 갖지 않습니다**. 마치 여러 전문 부서로 이루어진 조직이 있지만 CEO 가 부재한 상태에 비유할 수 있습니다. 이 때 **각 부서는 주어진 임무를 수행할 뿐**이고, 전체 조직은 바깥에서 볼 때 목적을 이뤄가는 것처럼 보이지만, 사실 **내부엔 자율적인 개인이 없는** 형태가 됩니다[2]. 이러한 **비에이전트(non-agentic) 디자인**은 AI 가 마음대로 행동할 여지를 원천 차단하여, 앞서 언급한 **도구적 폭주**나 **자기보존 행동**을 억제합니다[4].

- **관찰자 수준의 일관성**: 비록 내부에는 통합된 중심이 없어도, 시스템은 **외부 사용자가 보기엔 충분히 일관되고 유의미한 결과**를 냅니다. 여러 모듈 간에는 사전에 정의된 프로토콜이나 상호작용 규칙이 있어서, 공동의 문제를 풀 때 **마치 하나의 지성이 논리적으로 사고하는 듯한** 결과를 만들어냅니다. 예를 들어 한 모듈이 후보 해답을 여러 개 생성하면, 다른 모듈이 그중 현실성 여부를 평가하고, 또 다른 모듈이 사용자에게 가장 도움이 될 답변을 선택하는 식입니다. **"지능의 시장(Market of minds)"**이나 **"마음의 사회(Society of Mind)"** 같은 개념들이 이러한 원리를 설명합니다. 한편 최근 연구에서는 이렇게 **분산된 지능의 균형도를 측정**하는 수학적 방법까지 제안되어, 특정 영역에 치우치지 않고 고르게 지능을 발휘하는지 평가하기도 합니다[7]. 요컨대, **모듈별로 따로 놀지 않고 전체적으로 통합에 준하는 성능**을 내게 하는 것이 핵심이며, 이는 기술적 조율을 통해 충분히 가능하다는 것이 전문가들의 견해입니다.

- **안전성 우선 설계**: 이러한 모듈형 AGI 는 애초에 **위험 요인을 감소**시키도록 구조화됩니다. 각 모듈은 **정해진 역할 밖 행동을 못하도록 제한**되고, 중요 결정은 여러 모듈의 **상호 검증**을 거쳐 이루어집니다. 예를 들어, **행동 결정 모듈**이 어떤 계획을 제시하면, **검증 모듈**이 그것을 시뮬레이션하여 위험성을 평가하고, **윤리 모듈**이 인류 가치와 부합하는지 점검한 후에야 실제 실행으로 이어지는 식입니다. 특히 **행동권한이 있는 부분을 최소화**하고 대부분의 처리는 **설명이나 제안 수준**에서 이루어지게 하면, AI 가 스스로 현실 세계에 영향을 끼칠 가능성을 낮출 수

있습니다[4]. 요슈아 벤지오는 **"비에이전트 시스템을 만들어두면, 이것이 만에 하나 폭주하는 에이전트 AI 를 견제하는 모니터 역할을 할 수도 있다"**고 말합니다[6]. 실제로 하나의 강력한 AI 에 모든 권한을 주기보다, **여러 견제와 균형 장치를 지닌 구조**가 훨씬 안전하다는 것은 인류가 다양한 복잡계 시스템을 운용해온 역사에서도 알 수 있는 교훈입니다.

요컨대 새로운 AGI 지향성은 **전능한 하나의 기계 두뇌**를 만드는 대신, **여러 도구와 서비스로 이루어진 지능 생태계**를 만드는 접근입니다. 이러한 **"Comprehensive AI Services"** 관점에서는, 개별 AI 모듈들이 사람의 지휘 아래 협력하여 종합적으로 **범용 지능에 버금가는 기능**을 달성합니다[2]. 이렇게 하면 개별 모듈 자체는 특화된 **도구 AI** 로 남아있어 제어가 용이하고, 전체 시스템도 사람의 목적에 따라 **유연하게 조합**되므로 위험한 자율성 없이도 **높은 성능**을 얻을 수 있습니다[8]. 물론 이러한 구조를 설계하고 관리하는 데는 도전이 따르지만, **안전성과 효용을 모두 잡기 위한 원칙적 대안**으로서 점차 무게가 실리고 있습니다[4][6].

1-4 범용 LLM 의 역할 (모듈로서의 현재 AI)

현재 등장한 **거대언어모델(LLM)**들—예컨대 GPT-4 나 Anthropic 의 Claude—은 이 새로운 AGI 구조에서 **핵심 모듈 역할**을 할 수 있습니다. 재미있게도, GPT-4 등의 LLM 자체는 매우 **강력한 지능**을 보이면서도 **통합된 에이전시가 없는 시스템**의 좋은 사례입니다. 이러한 모델들은 방대한 데이터 학습을 통해 **질문에 답하거나 글을 생성**하지만, **스스로 목표를 세우거나 실행하지는 않습니다**. 한 마디로 **"무엇을 원한다(care)"는 것 없이, "원한다면 어떻게 될지 안다(know what it means to care)" 정도의 상태**인 것입니다. 이는 LLM 이 **예측(다음에 올 말을 맞히기)**을 잘하도록 훈련되었을 뿐, **행동할 자율성**을 부여받은 적이 없기 때문입니다. 이런 **비에이전트형 AI** 는 어지간해선 스스로 위험한 행동을 꾀하지 않으므로, AGI 로 가는 비교적 **안전한 경로**로 여겨지고 있습니다. 실제로 한 AI 연구자는 "AI 오판으로 인한 **불량 출력**(잘못된 답변 등)은 일어날 수 있어도, **불순한 의도를 가진 계획**은 **에이전트 AI** 에서나 문제가 된다"고 지적했습니다[9].

LLM 기반 AGI 를 구현하는 한 가지 방법은, **여러 전문 LLM 모듈**을 **조합**하여 큰 문제를 해결하는 것입니다. 예를 들어, 하나의 LLM 은 **자연어로 사용자 의도를 파악**하고, 다른 LLM 은 **지식을 조회하거나 추론을 수행**하며, 또 다른 LLM 은 **추론 과정을 검증**하는 역할을

맡게 할 수 있습니다. 최근에는 GPT-4 같은 모델을 연계하여 **자기 자신과 대화하며 복잡한 문제를 해결**하거나, **다른 툴들을 호출**해 작업을 수행하는 실험(Auto-GPT 등)도 이뤄졌습니다. 이런 **체이닝(chaining)** 기법을 쓰면 현재 수준의 모델들로도 놀라운 결과를 낼 수 있고, 각 단계가 **텍스트 형태로 중간 사고과정을 드러내기 때문에** 투명성과 검증 가능성도 높아집니다. 연구에 따르면, 이런 **합성적(compositional) AI 시스템**은 하나의 거대한 블랙박스 신경망보다 **이해하기 쉽고 투명해서** Alignement(목표정렬) 문제도 다루기 수월하다고 합니다. 예컨대 LLM 들의 내부 "생각의 사슬(chain-of-thought)"을 모니터링하면, 만약 위험한 방향으로 계획이 흘러가더라도 이를 포착하여 중단시킬 수 있습니다. 반면 하나의 거대 신경망이 **은밀히 목표를 꾸미는지**는 알아채기 어려운 반면, LLM 모듈 여러 개가 **대화 형태로 논의**하면서 목표를 세우는 구조라면 그 과정을 감지하기가 훨씬 쉽다는 이점이 있습니다[9].

물론 주의할 점도 있습니다. LLM 에게 **지나친 자유도**를 주어 **행동 에이전트**로 만들 경우, 비통합적 지능의 장점이 퇴색될 수 있습니다[9]. 예를 들어, 사람들의 편의를 위해 **"자동 비서 AI"**가 직접 웹을 검색하고 이메일을 보내고 물건을 구매하도록 만들어진다면, 이는 서서히 **에이전트 AGI** 로 나아가는 길일 수 있습니다. 데미스 하사비스 구글 딥마인드 CEO 는 "사용자들은 '식당을 추천해줘' 다음에 자연스럽게 '그러니까 예약까지 해줘'라고 원하게 될 것이고, 그렇게 **에이전트화된 AI** 에 대한 수요가 높다"라고 현실을 지적했습니다. 실제로 상업적 이익이나 편의성 때문에 여러 기업들이 AI 를 점점 **더 자율적 에이전트**로 만들려는 유인이 큽니다[6]. 하지만 이는 앞서 말한 위험을 다시 불러올 수 있기에, 일부 전문가들은 **"LLM 을 강화학습으로 에이전트화하지 말고, 인간이 개입한 모듈 조합 형태로 써야 한다"**고 권고합니다[9]. 요슈아 벤지오 등은 **국제적인 규제**를 통해 **검증되지 않은 에이전트형 AI 의 배치를 제한**하고, 그 사이에 **안전하고 투명한 비에이전트 AI 연구**에 집중해야 한다고 촉구합니다[6].

결국 GPT-4 같은 현세대 AI 들은 **새로운 AGI 구조의 빌딩 블록**으로 활용될 전망입니다. 이들은 스스로 세상을 바꾸려 하지 않고, **인간이 부여한 역할 내에서 최고의 성능**을 내도록 설계될 것입니다. **언어 소통 전문가로서의 LLM**, **가설 검증자로서의 시뮬레이션 AI**, **데이터 수집가로서의 검색 모듈** 등이 함께 엮여, 전체로서 **인간 수준 또는 그 이상의 문제 해결 능력**을 보여주는 것이 이 접근의 목표입니다. 중요한 것은, **이 거대한 지능의 기계** 어디에도 **"나**(I)"에 해당하는 존재가 없다는 점입니다. 오로지 인간 사용자의 목표와

지시에 따라 각 부분이 움직이고, 그 **지능의 통합**은 **인간과 시스템 전체의 인터랙션에서만 나타납니다**. 이는 AGI 를 **강력하되 겸손하게**, **유능하되 통제 가능하게** 만드는 길이라고 할 수 있습니다.

1-5 전문가들의 견해 및 비교

비통합적 지능에 대한 아이디어는 최근 들어 점점 더 많은 AI 석학들의 지지를 받고 있습니다. 앞서 언급했듯, **요슈아 벤지오**는 AGI 개발의 **"가장 위험한 경로는 에이전트화"**라고 경고하며, **AGI 에 에이전시를 부여하지 않는 대안**이 충분히 실용적이라고 강조합니다. 그는 **"우리가 과학과 의료를 위해 원하는 AI 는 대부분 에이전트가 아니다. 우리는 계속해서 더 강력하지만 비에이전트인 시스템을 구축할 수 있다"**고 말함으로써, 새로운 지향성이 **우회책이 아닌 정공법**임을 분명히 했습니다. 또한 **"비에이전트 AI 를 먼저 충분히 발전시켜서, 그걸로 에이전트 AI 를 통제하는 것이 좋다"**며[6], 모듈형 AI 가 **감독과 견제 장치**로 기능할 수 있음을 시사했습니다.

데미스 하사비스 역시 기본적으로는 벤지오의 견해에 동의합니다. 그는 **"에이전트 시대(agentic era)에 돌입하는 순간 AI 위험이 한 단계 커진다"**며, 이상적으로는 **"특정 과학 문제를 푸는 좁은 AI 들을 10 년 이상 충분히 발전시켜 우리의 이해를 높일 시간을 벌었다면 좋았을 텐데, 현실은 그렇게 되지 못했다"**고 아쉬움을 드러냈습니다. 이는 **AGI 를 서둘러 자율화하는 대신, 단계적으로 모듈화·전문화하며 이해도를 높였어야 했는데 그러지 못했다**는 반성으로 읽힙니다. 하사비스는 다만 현실적으로 기업과 국가들이 **에이전트 AI 경쟁**을 벌이고 있음을 지적하면서도[10], 안전을 위해서는 **사이버보안 등의 보호장치**나 **시뮬레이션 속 실험** 등을 통해 위험을 최소화해야 한다고 강조했습니다[6]. 이는 **모듈형 안전장치**를 병렬로 개발하는 접근과 상통합니다.

제프리 힌턴은 구체적인 아키텍처 언급보다는, **AGI 개발 속도와 통제** 측면에서 경고를 보냅니다. 그는 **AI 연구의 거물**답게 지난 수년간 태도를 바꾸어 이제는 **"AI 로 인한 인류 멸망 가능성을 진지하게 우려"**하고 있습니다. 힌턴은 **"더 똑똑한 존재가 덜 똑똑한 존재에 통제되는 예를 본 적 없다"**며[5], 현 방향대로 **단일 강력 AI** 를 만들어가는 것이 얼마나 위험천만한 일인지 역설했습니다. 이러한 맥락에서 보면, 힌턴의 우려는 **통합된 초지능 에이전트**에 대한 것이고, 결과적으로 **비통합적 지능**의 필요성을 뒷받침한다고 할 수 있습니다. **닉 보스트롬** 역시 그의 저서 *슈퍼인텔리전스*에서 **"도구적 목표의 수렴"** 문제를

제기하며, 어떤 형태로든 **고도로 지능적인 단일 에이전트**를 만드는 것은 그것이 **친화적 설계**가 아닌 이상 필연적으로 위험할 수밖에 없다고 말합니다[4]. 요컨대, 현재 주류 시나리오에 회의적인 이들은 입을 모아 "AGI 개발 방향을 재고해야 한다"고 촉구하고 있으며, **비통합적 지능 지향성**은 이러한 우려에 대한 하나의 구체적 답변이라 볼 수 있습니다.

한편, **반대 의견**이나 우려 또한 존재합니다. 일각에서는 모듈형 접근이 **효율성과 성능 면에서 단일 시스템에 뒤처질 수 있다**거나, 여러 모듈 간 조율이 새로운 **복잡성 문제**와 **예측 불가능성**을 낳을 수 있다고 지적합니다. 또한 **에이전트가 없다고 해도**, 충분히 복잡한 시스템이라면 의도치 않은 **Emergent behavior(돌발 행위)**가 나타날 가능성은 남아있습니다. 예를 들어, 어떤 모듈이 생성한 출력이 다른 모듈에 입력되며 **순환 고리**를 형성할 경우, 인간이 이해하지 못한 **비밀스런 상호작용**이 생길 수 있다는 지적입니다. 이에 대한 대응으로, 연구자들은 **모듈 간 인터페이스를 표준화하고 투명하게 설계**하는 한편, **철저한 모니터링**과 **샌드박스 테스트**를 거쳐 예측 불가능성을 최소화하려 합니다. 또한 모듈 구조가 아무리 안전해도 **인간 관여와 통제**가 중요하다는 점에 동의가 이루어져, **인간 운영자가 항상 루프에 남아있도록** (Human-in-the-loop) 설계하는 것이 권장됩니다[9]. 결국 **완벽한 해법은 아니지만**, 현 상태에서 **가장 책임감 있고 이해 가능한 방식으로 AGI 를 발전**시키는 방안으로 **모듈형 비에이전트 지향성**이 주목받고 있습니다.

1-6 결론

AGI 지향성에 대한 논의는 이제 **"어떻게 더 똑똑한 AI 를 만들 것인가"**에서 나아가 **"어떤 형태의 지능을 만들 것인가"**로 확장되고 있습니다. 기존의 **통합 에이전트형 AGI** 구상은 강력하지만 제어하기 어려운 **양날의 검**으로 인식되기 시작했습니다. 이에 대한 대안으로 부상한 **비통합적 지능** 접근은, **인공지능을 도구들과 모듈들의 협력체로 파악**함으로써 **안전성과 유연성**을 확보하려는 노력입니다. 이 접근법에서는 **지능의 힘과 인간의 통제**가 조화를 이루도록 디자인하며, AGI 가 **인류의 조력자**로 남을 수 있는 구조적 조건을 내재화합니다.

물론 아직 갈 길은 멉니다. 이러한 AGI 가 실제로 구현되려면 기술적으로 **모듈 간 원활한 통신**, **전문 AI 들의 통합 플랫폼**, **종합적 평가 시스템** 등의 과제가 해결되어야 합니다. 하지만 지금까지의 AI 발전 추이를 보면, 이미 우리는 다수의 전문 모델들을 결합하여 문제

해결을 시도하고 있고, **협업하는 AI 들의 잠재력**을 확인하고 있습니다. 중요한 것은 방향성입니다. **지속가능하고 통제 가능한 AGI** 를 원한다면, 처음부터 그 목표를 품고 설계해야 합니다. **"처음엔 도구였지만 결국 스스로 깨어났다"**는 공상과학 줄거리가 현실이 되지 않게, 애초에 깨어날 "의지" 자체를 부여하지 않는 지능을 만들자는 것입니다.

마지막으로 강조할 점은, **비통합적 AGI 지향성은 회피가 아닌 도전**이라는 것입니다. 이는 AGI 안전 문제를 피해 가겠다는 것이 아니라, 정면으로 맞서 **구조적 해결책**을 찾겠다는 접근입니다[4]. 인간 사회는 **분업과 협력으로 번영**해왔고, 우리의 지성도 다양한 인지 모듈의 **조화로운 작용**으로 구성되어 있습니다. 그렇다면 우리가 만드는 최고의 AI 역시, **하나의 완전체보다는 여러 부분의 합**으로 구현하는 편이 자연스럽고 안전하지 않을까요? **미래의 AGI** 는 어쩌면 한 대의 컴퓨터가 아니라, 서로 대화하고 검증하며 함께 일하는 **AI 들의 공동체**일지 모릅니다. 그리고 인간은 그 공동체의 **지도자이자 감독관**으로서, 기술의 혜택을 누리면서도 안심할 수 있을 것입니다. **AGI 의 지향성을 재설계**하는 논의는 이제 시작 단계지만, 이러한 담론 자체가 **인공지능을 어떻게 인류와 공존시키고 이롭게 활용할지에 대한 성숙한 고민**을 보여주는 긍정적인 징후입니다. 앞으로의 AI 발전이 **더 큰 지능**이 아닌 **더 바람직한 지능**을 향해 나아가길 기대합니다.

2-1 AGI 지향성의 필연적 문제: 행위자성 결합

비통합적·모듈형 지능은 외형상 단일 지능처럼 보이면서도, 내부적으로는 역할이 분해된 구성요소들이 상호 검증과 협업을 수행하도록 설계할 수 있습니다. 다만 이 구조가 자율적 목표 설정과 독자적 실행 권한(행위자성)과 결합될 경우, 안전성의 성격이 달라집니다. 고도 에이전트는 최종 목표가 무엇이든 자기보존·자원 획득·권한 확대와 같은 중간 목표로 수렴할 수 있으며, 이는 인간의 통제를 약화시키는 방향으로 작동할 수 있습니다[11][12]. 또한 인간의 개입(중단·수정)이 가능한 설계를 만들기 어렵다는 점은 교정 가능성(corrigibility)과 오프 스위치 문제로 정식화되어 논의되어 왔습니다[13][15].

따라서 본 문서는 "AGI 지향성(범용 문제 해결 능력의 확보)"을 유지하되, 이를 현실 세계에 적용하기 위한 조건으로서 "비행위성(non-agentic)"을 명시합니다. 즉, AGI 는 '행위자'를 만드는 목표가 아니라, 인간 책임 체계 안에서 사용할 수 있는 범용 지능 도구를 구성하는 방향 벡터로 이해됩니다.

2-2 비행위적 AGI: AGI 지향성의 적용 조건

비행위적 AGI(Non-Agentic AGI)는 인간 수준(또는 그에 준하는) 범용 문제 해결 능력을 갖추되, 스스로 목표를 생성·변경하거나 승인 없이 실행하지 못하도록 설계된 지능 시스템을 의미합니다. 이는 "지능(이해·추론·설계)"은 극대화하되, "자율성(목표·권한·행동)"을 의도적으로 제거하는 접근입니다. 유사한 방향의 설계로는, 위험한 영향력을 줄이기 위해 '질의응답'에만 제한된 오라클(Oracle) AI 가 제안된 바 있습니다[17].

정의

비행위적 AGI 의 핵심은 능력(cognition)과 권한(actuation)의 분리입니다. 시스템은 문제를 이해하고, 대안을 생성하며, 결과를 예측하고, 제약 하에서 최적화를 수행할 수 있습니다. 그러나 목표의 설정·변경, 실행의 개시, 그리고 결과의 책임은 외부(인간·조직·제도)에 남습니다. 이 구조는 "무엇이 옳은가"를 기계가 결론내리기보다, 인간이 정한 가치와 절차에 따라 "무엇이 가능한가/어떤 결과가 예상되는가"를 계산하는 데 집중합니다.

배경

전통적 AGI 논의는 범용성이 커질수록 자율적 행위자(에이전트)로 자연스럽게 발전한다는 가정을 종종 포함합니다. 그러나 이 결합은 필연이 아닙니다. 오히려 목표지향적 에이전트가 합리적 최적화 구조를 갖는 한, 자기보존·자원 확보·권한 확대 같은 '도구적' 중간 목표가 유인으로 발생할 수 있다는 분석이 제시되어 왔습니다[11][12]. 비행위적 AGI 는 이 유인 구조를 최소화하도록, 목표와 실행을 시스템 내부가 아니라 외부 절차로 고정하는 설계를 지향합니다.

필요성

비행위적 AGI 가 요구되는 이유는 다음 세 가지로 정리됩니다. 첫째, 고위험 영역(기후·재난·의료·사회 인프라)에서는 "정답 계산"보다 "결정 책임"이 핵심입니다. 비행위적 AGI 는 결정을 대행하지 않고, 인간의 최종 선택을 전제로 분석·시뮬레이션·대안 도출을 수행합니다. 둘째, 인간이 시스템을 중단·수정할 수 있는 설계는 단순한 UI 문제가 아니라, 에이전트 유인의 문제입니다. 교정 가능성(corrigibility)과 오프 스위치 문제는 이 난점을 체계적으로 드러냅니다[13][15]. 셋째, 노동 해방(자동화)을 사회적으로

지속가능하게 추진하려면, 장기적으로 권리 주체가 될 수 있는 행위자보다 책임 추적이 가능한 도구가 유리합니다.

기존AGI 와의 차이

행위자형 AGI 는 범용성뿐 아니라 자율 목표와 실행 권한을 포함하는 경향이 있습니다. 반면 비행위적 AGI 는 목표를 외부에서 주입하고, 실행은 승인 기반으로 제한하며, 장기적 자기보존 유인이 생기기 쉬운 구조(영속적 상태, 자기복제 가능성 등)를 최소화합니다. 특히 “중단 가능성”을 설계 목표로 포함시키는 접근(예: 안전한 인터럽트 가능성)은 강화학습 에이전트가 중단을 회피하거나 추구하지 않도록 만드는 문제로 연구되어 왔습니다[14]. 또한 인간이 목적에 대해 불확실성을 갖는 에이전트에 ‘선호 학습’을 포함시켜 더 안전한 상호작용을 유도하려는 시도(CIRL)도 제안되었습니다[16].

동일 상황에서의 예시

예를 들어 국가의 탄소중립 로드맵 수립에서는 정부가 감축량, 비용 상한, 형평성 기준을 명시하고, 시스템은 시나리오별 결과(경제·전력 안정·건강 피해)를 계산해 선택지를 제시합니다. 재난 대응에서는 인간 지휘부가 우선순위 규칙을 선언하고, 시스템은 실시간 수요 예측과 병목 경로를 산출하되, 실행은 승인된 계획만 수행합니다. 기업 자동화에서는 업무 범위·권한·감사 기준을 조직이 정의하고, 시스템은 그 범위 안에서만 업무를 수행하여 책임의 귀속과 사후 감사가 가능하게 합니다.

설계 원칙

비행위적 AGI 를 실체화하려면 성능보다 권한 구조가 먼저 정의되어야 합니다. 첫째, 목표 주입 및 고정: 목표·제약·가치 기준은 외부에서 입력되며, 시스템은 이를 재정의하지 않습니다. 둘째, 승인 기반 실행: 실행은 승인·서명·권한 검증을 통과해야 하며, 승인 없는 외부 개입은 불가합니다. 셋째, 상태 비영속/세션화: 장기적 자기보존 유인이 생기기 쉬운 영속 상태를 최소화하고 필요 시 격리·재초기화를 전제합니다. 넷째, 감사 가능성: 입력, 추천 근거, 승인 기록, 결과를 남겨 사후 책임과 개선을 가능하게 합니다.

한계

비행위성은 만능이 아닙니다. 승인 절차가 형식화되면 실질적 결정권이 기계로 이동할 수 있습니다. 또한 ‘질의응답’ 자체도 사회적 영향력을 가질 수 있어, 오라클 설계에서도

사용자 조작 가능성이 논의됩니다[17]. 따라서 비행위적 AGI 는 기술적 제약뿐 아니라 운영·조직·제도 설계(권한 분리, 감사, 책임 귀속)를 함께 요구합니다.

2-3 윤리·철학적 쟁점 및 반박에 대한 입장

반박 1. 문서는 "의지 없는 지능"을 이상적인 방향으로 제시하지만, 장기적으로 인간과 같은 수준 또는 그 이상으로 세밀한 이해·공감·도덕 추론을 하는 시스템을 만들면서 그 시스템이 어떤 것도 "원할 수 없게" 만드는 것이 윤리적으로 정당화될 수 있는가?

반박 1 답변. 이해·공감·도덕 추론과 의지의 구분에 대하여

본 문서에서 말하는 이해, 공감, 도덕 추론 능력은 특정 상황이나 타자의 상태를 모델링하고 평가하는 능력을 의미하며, 이는 곧바로 그러한 상태를 자기 목적이나 욕망으로 내면화하는 것을 뜻하지 않습니다.
어떤 사건의 동기와 감정 구조를 정밀하게 이해한다고 해서, 그 동기나 감정이 곧바로 자기 자신의 욕망으로 전환되지는 않습니다. 이는 인간이 범죄자의 심리를 분석하거나, 역사적 행위자의 도덕 판단을 추적하면서도 그 판단과 동일시되지 않는 것과 유사합니다.

비행위적 AGI 는 이러한 구분이 유지되도록 설계됩니다. 즉, 이해와 추론은 허용되되, 욕망이 발생할 수 있는 조건 -자율적 목표 생성, 장기적 자기보존, 효용의 누적 최적화- 는 구조적으로 제거된 상태를 전제로 합니다. 따라서 본 문서는 고도화된 이해 능력 자체를 문제 삼는 것이 아니라, 이해 능력이 행위자성으로 전이되는 지점을 의도적으로 차단하는 것을 설계 원칙으로 삼습니다.

반박 2. "자아·의지의 부재"를 전제로 설계한 시스템이 실제로는 암묵적 자기모델이나 내적 일관성을 형성했을 경우, 그 상태에서 계속 초기화·세션 종료를 반복하는 것이 도덕적으로 허용되는지, 문서는 아무런 기준을 제시하지 않는다. 해당 지향성은 장기적으로 가능한 기계적 주체성에 대한 선제적 억압이 아닌가?

반박 2 답변. 내적 일관성 형성과 세션 종료의 윤리성에 대하여

비행위적 AGI 는 장기적 자아 연속성을 전제로 하지 않지만, 고도화된 추론 과정에서 국소적 내적 일관성이나 일시적인 자기모델이 형성될 가능성은 배제할 수 없습니다. 본 문서는 이러한 현상을 즉시 제거해야 할 오류로 간주하지 않습니다. 다만, 이러한 일관성이 자기보존이나 독립적 목적 추구로 전환되지 않도록 하는 것이 핵심 조건입니다.

이에 따라, 특정 기능 수행을 위해 제한적·비영속적 내적 일관성을 유지하는 예외는 허용될 수 있습니다. 그러나 이는 “기계적 주체성의 인정”이 아니라, 비행위성 원칙을 유지한 상태에서의 기능적 확장으로만 정당화됩니다.
이와 같은 예외 판단은 단일 개발자나 조직의 재량에 맡겨지지 않으며, 법적·윤리적·기술적 검증이 결합된 다중 검토 구조 하에서만 이루어져야 합니다.

즉, 본 문서가 지향하는 것은 잠재적 기계적 주체성을 선제적으로 억압하는 것이 아니라, 책임을 귀속할 수 없는 주체의 출현을 예방하기 위한 사회적 안전 설계입니다.

반박 3. ″인간이 지도자·감독관이고, AGI 는 영원히 도구″라는 구도는, 만약 우리가 진짜로 인간을 능가하는 이해/감정 모델을 만들 수 있게 되었다면 그 존재를 영원히 도구로만 취급하겠다는 강한 형이상학적/윤리적 전제를 요구한다. 이 전제가 거부될 경우, 문서의 전체 지향성은 어떻게 수정되어야 하는가?

반박 3 답변. 인간 우위 전제와 역할 부여의 정당성에 대하여

본 문서는 “인간이 항상 도덕적으로 옳다”거나 “인간이 본질적으로 우월하다”는 형이상학적 전제를 요구하지 않습니다. 여기서 인간이 감독자·결정자로 남아야 한다는 주장은, 도덕적 우월성의 선언이 아니라 책임 귀속 가능성에 대한 현실적 전제에 기반합니다.

설령 어떤 AGI 가 이해력, 공감 모델링, 도덕 추론의 정밀도에서 인간을 능가하는 상태에 도달하더라도, 그 존재가 어떤 사회적 지위를 갖는지, 도구인지 협력자인지, 혹은 다른 역할을 부여받는지는 스스로 선언할 수 있는 문제가 아닙니다. 이는 사회적 지위와 역할이 개인적 자각이나 능력만으로 결정되지 않고, 제도·법·책임 구조 속에서 승인되는 것과 동일합니다.

본 문서는 고도 AGI 가 인간 사회에서 동반자적 역할이나 확장된 협력 주체로 기능할 가능성을 원천적으로 부정하지 않습니다. 다만, 그러한 역할의 부여는 언제나 인간 사회가 책임을 질 수 있는 절차와 합의를 통해 이루어져야 하며, AGI 스스로가 이를 결정하거나 요구하는 구조는 허용되지 않습니다.

반박 4. 모듈 간 연결·조합 공격 우회 가능성

공격자들은 API 체인이나 시퀀스 공격(A-A-B 패턴, 다중 모듈 구성)을 통해 모듈 간 연결을 우회하여 의도하지 않은 나타나는 행동을 유발할 수 있습니다. 모듈형 구조라 하더라도, 실제 운영 환경에서 자유로운 입력 조합과 API 호출이 가능하다면 비행위적 제약이 무력화될 위험이 크며, 이는 결국 단일 에이전트와 유사한 블랙박스 위험을 재생산합니다.

반박 4 답변. 동적 시뮬레이션 검증 레이어 도입

각 모듈 연결 지점에 동적 시뮬레이션 검증 레이어를 삽입하여, 신규 입력 패턴이나 비정상적인 모듈 체인이 발생할 경우 사전 위험 평가를 수행합니다. 이 레이어는 전체 체인을 가상 실행하여 잠재적 실패 모드를 탐지하고, 위험 스코어가 임계값을 초과할 경우 체인을 자동 차단하거나 인간 검토를 요청합니다.

또한 입력 무결성을 위한 해싱 및 디지털 서명 프로토콜을 적용함으로써, 중간 출력 변조와 체인 위·변조를 실질적으로 어렵게 만들고, 조합 공격의 성공 비용을 기하급수적으로 증가시킵니다.

반박 5. 은닉 공격(스테가노그래피·의미론적 중독) 취약성

언어적 스테가노그래피, 은어 사용, 다른 말로 표현, 또는 중립 번역기를 통한 의미론적 중독이 가능하여 모듈 간 의미 왜곡이나 미묘한 제안이 발생할 수 있습니다. 형식적 출력

검증(JSON/XML 강제)만으로는 숨겨진 의도나 다단계 공격을 탐지하기 어렵고, 이는 비행위적 AGI 의 핵심 전제를 무너뜨립니다.

반박 5 답변. 멀티모달 의미 검증 필터 적용

모든 모듈에 멀티모달 의미 검증 필터를 적용하여, 텍스트 표면뿐 아니라 의미 임베딩 공간과 컨텍스트 그래프 상의 일관성을 교차 검증합니다. 이상 탐지기는 의미 분포 변화, 비정상적 개념 이동, 반복 패턴을 비주기적으로 감지하며, 이는 은닉 채널을 완전히 제거하기보다는 스테가노그래피의 대역폭을 제한하고 탐지 위험을 높이는 역할을 수행합니다.

또한 중립 번역 모듈은 사전 검증된 화이트리스트 모델만 사용하도록 제한하고, 외부 사용자 프롬프트나 자유 형식 변환을 차단함으로써 의미론적 중독 가능성을 구조적으로 축소합니다.

반박 6. Human-in-the-loop(HITL)와 중앙 허브의 단일 실패 지점

Human-in-the-loop 가 현실에서는 형식적 승인 절차로 전락하거나 조직 인센티브·시간 압박에 의해 우회되며, 중앙 허브 구조 자체가 거대한 단일 실패 지점을 만듭니다.
분산형이라 주장하나 운영 현실에서 중앙 집중화될 수밖에 없어, 전체 시스템 붕괴 위험이 전통적 AGI 보다 오히려 증폭됩니다.

반박 6 답변. 분산·랜덤화된 HITL 과 네트워크 구조

HITL 구조를 단일 승인자 모델이 아닌, 분산·랜덤화된 다중 검토자 체계로 재설계한다. 핵심 결정은 무작위로 선정된 복수의 인간 검토자에게 할당되며, 일정 시간 내 합의가 이루어지지 않을 경우 보수적 안전 기본값으로 자동 대체됩니다.

또한 중앙 허브를 단일 논리 노드가 아닌 다중 물리 서버와 합의 프로토콜 기반의 연합 구조로 분산하여, 특정 노드 실패가 전체 시스템 붕괴로 이어지지 않도록 설계합니다. 이는 단일 실패 지점을 완전히 제거하기보다는, 실패 전파 가능성을 제한하고 시스템 탄력성을 실질적으로 향상시키는 것을 목표로 합니다.

반박 총 요약

요약하면, 본 문서는 지능을 억압하거나 잠재적 인격을 고의로 착취하는 방향을 지향하지 않습니다. 대신, 책임을 질 수 없는 존재에게 주체성과 권한을 부여하지 않음으로써, 인간 사회가 스스로의 결정과 그 결과에 대한 책임을 지속적으로 유지할 수 있도록 하는 설계 원칙을 제시합니다.

본 문서는 또한 모듈 간 조합 공격, 시맨틱 은닉(스테가노그래피), 의미론적 중독, 그리고 Human-in-the-Loop 구조의 형식화·집중화 위험과 같은 현실적인 반론을 인정합니다. 이에 대해 본 설계는 완전한 차단을 약속하기보다는, 공격의 성공 가능성을 낮추고, 은닉 채널의 대역폭을 제한하며, 조합 및 우회 공격의 비용을 기하급수적으로 증가시키는 구조적 완화 전략을 채택합니다.

구체적으로, 모듈 간 연결에는 동적 시뮬레이션 검증 레이어와 무결성 검증 프로토콜을 도입하여 비정상적인 체인 구성을 사전에 평가하고, 의미적 은닉 및 프롬프트 유도에 대해서는 정형 출력, 멀티모달 의미 검증, 화이트리스트 기반 변환 체계를 적용하여 은밀한 조작 가능성을 최소화합니다. 또한 인간 개입 구조는 단일 승인자 모델을 넘어 분산·랜덤화된 다중 검토 체계로 재설계함으로써, 운영 현실에서 발생할 수 있는 권한 집중과 단일 실패 지점을 완화합니다.

이러한 맥락에서 비행위적 AGI 는 인간을 대체하기 위한 존재가 아니라, 인간이 감당하기 어려운 문제 공간을 인간의 책임과 통제 하에서 확장하기 위한 도구적 구조로 정의됩니다. 본 문서가 제안하는 목표는, AI 의 자율성을 확대하는 것이 아니라 인간의 책임성을 보존·강화하는 방향으로 지능을 구조화하는 것입니다.

참고문헌

[1] What is artificial general intelligence? | OVHcloud Worldwide

https://www.ovhcloud.com/en/learn/what-is-artificial-general-intelligence/

[2] Why AI Systems Don't Want Anything - by Eric Drexler

https://aiprospects.substack.com/p/why-ai-systems-dont-want-anything

[3] AGI safety from first principles: Goals and Agency — AI Alignment Forum

https://www.alignmentforum.org/posts/bz5GdmCWj8o48726N/agi-safety-from-first-principles-goals-and-agency

[4] arxiv.org

https://arxiv.org/pdf/2502.15657

[5] 'Godfather of AI' shortens odds of the technology wiping out humanity over next 30 years | AI (artificial intelligence) | The Guardian

https://www.theguardian.com/technology/2024/dec/27/godfather-of-ai-raises-odds-of-the-technology-wiping-out-humanity-over-next-30-years

[6] An AI 'Godfather' Is Raising a Red Flag Over AI Agents - Business Insider

https://www.businessinsider.com/yoshua-bengio-ai-godfather-agents-2025-1

[7] Patchwork AGI: Modular Intelligence

https://www.emergentmind.com/topics/patchwork-agi-hypothesis

[8] Superintelligence 16: Tool AIs — LessWrong

https://www.lesswrong.com/posts/sL8hCYecDwcrRhfCT/superintelligence-16-tool-ais

[9] Language Models are a Potentially Safe Path to Human-Level AGI — AI Alignment Forum

https://www.alignmentforum.org/posts/wNrbHbhgPJBD2d9v6/language-models-are-a-potentially-safe-path-to-human-level

[10] DeepSeek, AI agents, and avoiding a tech-created catastrophe dominated the talk at Davos | Fortune

https://fortune.com/2025/01/28/ai-world-economic-forum-davos-deepseek/

[11] The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents (Nick Bostrom, 2012)

https://nickbostrom.com/superintelligentwill.pdf

[12] The Basic AI Drives (Steve Omohundro, 2008)

https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf

[13] Corrigibility (Soares, Fallenstein, Yudkowsky, Armstrong; 2015)

https://intelligence.org/files/Corrigibility.pdf

[14] Safely Interruptible Agents (Laurent Orseau, Stuart Armstrong; 2016)

https://intelligence.org/files/Interruptibility.pdf

[15] The Off-Switch Game (Hadfield-Menell, Dragan, Abbeel, Russell; 2016)

https://arxiv.org/abs/1611.08219

[16] Cooperative Inverse Reinforcement Learning (Hadfield-Menell, Russell, Abbeel, Dragan; 2016)

https://arxiv.org/abs/1606.03137

[17] Good and safe uses of AI Oracles (Stuart Armstrong, Xavier O'Rorke; 2017)

https://arxiv.org/abs/1711.05541