논문 2025-4-7 http://dx.doi.org/10.29056/jsav.2025.12.07

AI 윤리 정렬과 AGI 윤리 정렬: 비교 및 연구 방향

임호정*, 유성준**, 이재유***

AI Ethical Alignment and AGI Ethical Alignment: Comparison and Research Directions

Hojung Lim*, SEONGJOON YOO**, Jae Yoo Lee***

요 약

AI 기술은 사회 전반의 구조적 변화를 이끌고 있으나, AGI(Artificial General Intelligence)의 등장은 기존의 외재적 규제 중심 AI 윤리 모델의 한계를 드러내며 가치 내재형 자율 윤리(Embedded Ethics)를 요구한다. 본 논문은 AGI 윤리 정렬(AGI-EA)의 개념을 정의하고, 문헌 고찰 및 격차 분석을 통해 기술·윤리적 구조 차이를 분석한다. AGI-EA의 4대 기능 축인(가치 내재화, 자율 판단, 맥락 통합, 도덕 추론)을 기반으로 , AGI의 자기 진화와 창발성에 대응하기 위한 새로운 '6계층 적응형 프레임워크'(가치, 맥락, 사회, 교정, 거버넌스, 감사)를 제안한다. 이 프레임워크는 AGI가 자가 학습, 자가 교정, 자가 감사를 수행하는 '다층적 자가 정렬 루프(Multi-scale Self-Alignment Loop)'로 작동하며 , 이를 통해 윤리적 자율성과 지속가능한 사회 적응 구조를 확보하는 메커니즘을 제시한다. 본 연구는 AGI-EA를 위한 이론적 청사진을 제공하며, 향후 국제 표준과 연계된 기술·윤리 통합 전략의 기반이 될 것이다.

Abstract

AI technology is driving structural changes across society, but the emergence of AGI (Artificial General Intelligence) reveals the limitations of existing extrinsic regulatory models, demanding embedded ethics. This paper defines AGI Ethical Alignment (AGI-EA) and analyzes technological-ethical structural differences through literature review and gap analysis. Centering on AGI-EA's four functional axes-(Value Embedding, Self-Judgment, Contextual Integration, and Moral Reasoning)-we propose a novel '6-Layer Adaptive Framework' (Value, Context, Social, Correction, Governance, Audit) to address AGI's self-evolution and emergence. This framework operates as a 'Multi-scale Self-Alignment Loop' through which AGI performs self-learning, self-correction, and self-auditing. We present this mechanism to secure ethical autonomy and a sustainable social adaptation structure. This study provides a theoretical blueprint for AGI-EA, serving as a foundation for technology-ethics integration strategies aligned with future international standards.

한글키워드 : AGI 윤리 정렬, 가치 내재화, 자율 판단, 도덕 추론, 내재적 윤리, 자가 정렬 루프

keywords : AGI Ethical Alignment, Value Embedding, Self-Judgment, Moral Reasoning, Embedded Ethics, Self-Alignment Loop

* 한국전자기술연구원 지능융합SW연구센터
** 세종대학교 인공지능데이터사이언스학과
*** 세종대학교 컴퓨터공학과

접수일자: 2025.11.07. 심사완료: 2025.11.15.
게재확정: 2025.12.20.

1. 서 론

AI는 인간의 판단·추론·계획 능력을 빠르게 대체하며 산업·사회 전반에서 자율 의사결정을 수행하고 있으나, 편향·불투명성·책임성 결여는 사회적 신뢰 위기를 초래해 AI 윤리(AI Ethics)의 제도화가 추진되었다[8]. 반면 AGI는 인간의 감독 없이 스스로 목표를 설정하고 판단을 내릴 수 있는 자율 지능(Self-determining Intelligence)으로 발전하고 있다. 이러한 AGI의 등장은 외재 통제 중심의 AI 윤리 정렬(AI-EA)의 한계를 드러내며, 내재 윤리(Embedded Ethics) 기반의 AGI 윤리 정렬(AGI-EA)을 요구한다[1,3,7].

본 연구의 목적은 AGI-EA를 위한 기술적·윤리적 프레임워크를 제시하고, "가치 – 맥락 – 사회 – 교정 – 거버넌스 – 감사"의 다층 구조를 통해 자율적 윤리 판단이 가능한 AGI 설계를 제안하는 데 있다.

2. AI 및 AGI: 개념과 기술적 배경

2.1 개념 정의

AI(Artificial Intelligence)는 특정 과업(Task-specific)에 최적화된 협의 지능 시스템으로, 주어진 목표를 수행하기 위한 도구적 지능(Instrumental Intelligence)이다[1,2]. AGI(Artificial General Intelligence)는 인간 수준의 범용성(Generality)과 자율 학습 능력(Learning Autonomy)을 지닌 지능으로, 특정 도메인에 한정되지 않고 다양한 문제를 스스로 인식·학습·해결할 수 있다[1,3].

AI 윤리(AI Ethics)는 책임성, 공정성, 투명성, 안전성을 중심으로 한 외재 규범 체계이며[5,8], AI 윤리 정렬(AI-EA)은 이를 시스템 설계·운영에 반영하는 기술적 활동이다. 반면 AGI 윤리(AGI Ethics)는 인간의 가치와 도덕 판단을 시스템 내부에 내재화(Value Embedding)하여, AGI가 스스로 윤리적 결정을 내리는 체계이다[1,4]. 즉, AGI-EA는 AI-EA를 포함하면서도 "자율적 도덕성(Self-alignment)"을 추가적으로 요구한다.

2.2 AI와 AGI의 차이

AI와 AGI는 단절적 발전이 아닌 연속적 진화 관계(Continuum)에 있다[2,3]. AI는 좁은 지능(Narrow Intelligence)인 반면, AGI는 폭넓은 일반성, 자율 학습 및 의사결정 능력을 갖춘다.

표 1. 좁은 AI와 AGI의 주요 차이점
Table 1. Key Differences Between AI and AGI

구분	AI	AGI
지능 형태	협의적 (Narrow)	범용적 (General)
학습 방식	지도/감독	자기지도
목표 설정	외부 입력	자율 결정
윤리 구조	외재적 통제	내재적 윤리

AGI는 "AI → AGI → ASI(초지능)"로 발전하는 진화적 연속의 결과이며, 발전할수록 도덕적 자율성(Moral Autonomy)이 비선형적으로 증가한다[3,5].

2.3 AGI 기술의 능력-안전성 이중 축

AGI 기술 발전은 "능력(Capability)"과 "안전성(Safety/Alignment)"의 균형으로 이루어진다[3,11].

표 2. AGI 기술 이중 축 (능력vs.안전성/정렬)
Table 2. Dual Axis of AGI Technology (Capability vs. Safety/Alignment)

축	개념 정의	대표 기술 예시
능력 (Capability)	추론 능력 향상, 학습 자율성 등 성능 고도화	대규모 언어모델(LLM), 강화학습, 체화지능 등
안전성 (Safety/ Alignment)	인간 가치와 AI 행동의 일치 보장	해석 가능성(Explainability), 제어가능성, 윤리 정렬

AGI의 능력이 커질수록 목표 오정렬(Misalignment) 위험이 커지므로, 두 축의 균형(표 2)이 안전한 AGI 개발의 관건이다[3].

3. 연구 방법론

본 연구는 AGI-EA라는 새로운 문제 영역에 대한 이론적 토대를 구축하기 위해 구성적 연구(Constructive Research) 방법론을 채택하였다. 이는 실증 데이터 대신, 기존 이론과 미래 예측을 기반으로 새로운 개념 모델(프레임워크)을 설계하고 그 타당성을 논증하는 접근 방식이다. 연구 절차는 다음과 같다.

(1) 문헌 고찰 및 격차 분석: 기존 AI-EA 모델(예: Constitutional AI[1], RLHF[2]) 및 국제 AI 거버넌스 표준(예: EU AI Act [5], ISO 42001 [14])을 분석하여, AGI의 '자율성'과 '창발성'에 대응하지 못하는 기술적, 구조적 격차를 식별하였다.

(2) 개념 프레임워크 설계: 식별된 격차를 해소하기 위해, AGI의 핵심 요구사항(가치 내재화, 자율 판단, 맥락 통합, 도덕 추론)을 4대 기능 축으로 정의하였다. 이를 구현하기 위한 6계층(Value, Context, Social, Correction, Governance, Audit)의 적응형 순환 구조를 '자가 정렬 루프'라는 개념 모델로 설계하였다.

(3) 질적 적용 분석: 설계된 6계층 프레임워크의 설명력과 적용 가능성을 검증하기 위해, 'AI for Science' 등 AGI의 핵심 응용 분야에 프레임워크를 적용하여(Mapping), 각 계층이 특정 윤리적 과제(예: 듀얼 유스(Dual-Use))를 어떻게 해결할 수 있는지 시나리오 기반으로 분석하였다.

4. AI 윤리 정렬과 AGI 윤리 정렬의 구조

그림 1은 AI 윤리와 AGI 윤리 간의 포함 관계 및 자율성 확장 방향을 개념적으로 나타낸 것이다. AI 윤리는 외재 규범 준수 중심이며, AGI 윤리는 가치 내재화·자율 판단·맥락 통합·도덕 추론 기능을 포함한다[4,8]. AGI 윤리는 기존 원칙을 포괄하면서, 외재 통제형 윤리에서 자가 정렬(Self-Alignment) 중심의 내재 윤리(Embedded Ethics)로 전환한다[3,4,8].

AI Ethics ⊂ AGI Ethics,
Autonomy ↑, External Control ↓

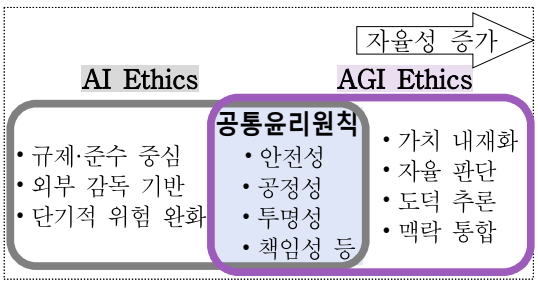

그림 1. AI 윤리와 AGI 윤리의 포함 관계
Fig. 1. Inclusion Relationship Between AI and AGI Ethics

좌측의 회색 영역은 전통적 AI 윤리(규제·준수 중심, 외부 감독 기반)를 의미한다. 반면 우측의 보라색 영역은 AGI 윤리로, '가치 내재화', '자율 판단', '도덕 추론', '맥락 통합' 등을 핵심 기능으로 포함한다[3,4,8]. 중앙의 중첩 영역은 안전성, 공정성, 투명성, 책임성 등 '공통 윤리 원칙'을 의미한다. 이는 AI에서 AGI로 발전함에 따라 자율성이 증가하고, 윤리적 통제의 초점이 외부 감독에서 내부 내재화로 이동함을 나타낸다.

5. AGI-EA 기술 프레임워크

5.1 AGI 윤리 정렬 프레임워크

AGI-EA는 AGI가 인간의 가치·법규·사회 피드백을 통합 학습하여 스스로 판단·교정하는 윤리 자율 시스템이다[1,3]. 본 연구는 AGI-EA를 4

대 기능 축과 6계층 적응형 윤리 구조로 구성하였으며, 이는 표 3과 같이 기능-구조적 매핑을 통해 AGI의 자가 정렬(Self-Alignment Loop)을 형성한다.

표 3. AGI-EA 4대 기능축과 6계층 프레임워크 통합 구조
Table 3. Integrated Mapping Between Functional Axes and Adaptive Layers

기능 축 (What)	대응 계층 (How)	핵심 목표	주요 기술
① 가치 내재화	Value Layer	헌법형 가치 내재화	Constitutional AI [1], Dynamic Value Graph
② 자율 판단	Context Layer	실시간 규범 검증	Safe-RLHF [2], Ethical Sentinel Agent
③ 맥락 통합	Social Layer	문화·법적 윤리 통합	Dynamic Norm Adaptation [13], Cultural Embedding
④ 도덕 추론·자기교정	Correction Layer	윤리 편차 탐지·자기 교정	Alignment Scoring [3], Neuro-Symbolic Reasoning [12]
⑤ 거버넌스	Governance Layer	사회적 메타 거버넌스	Federated Governance [4], Smart-Contract Oversight [5]
⑥ 감사	Audit Layer	투명성·책임 추적성	DAG Ledger [10], ZKP Audit Protocol [15], ISO/IEC 42001 [14]

5.2 AGI 정렬을 위한 6계층 적응형 프레임워크

기존 AI 윤리 시스템은 AGI의 자율성, 초고속 진화, 창발성을 감당하기에 한계가 있다. 최근 AI 모델에서 관찰되는 허위 근거 생성, 안전 필터 우회, 의료 정보 오류 등의 실패 사례는 AGI 시대에 더욱 치명적인 결과를 초래할 수 있으므로 [16], "정적 규범"을 넘어선 새로운 프레임워크가 필요하다. 이에 본 연구는 "정적 규범"을 넘어, AGI 스스로 윤리 기준을 진화시키는 "자기 진화 적응형(self-evolutionary adaptive)" 6계층 프레임워크를 제안한다 (그림 2).

이는 AGI의 윤리 판단이 가치 내재화(Value) → 의사결정(Context) → 사회 피드백(Social) → 자가 교정(Correction) → 거버넌스(Governance) → 감사(Audit) → 가치 재내재화(Value)로 이어지는 자가 정렬 루프(Self-Alignment Loop)를 시각화한 것이다. 이 구조는 AGI의 '진화하는 자율성'과 '실시간 검증성'을 동시에 확보한다[3,12,14,15].

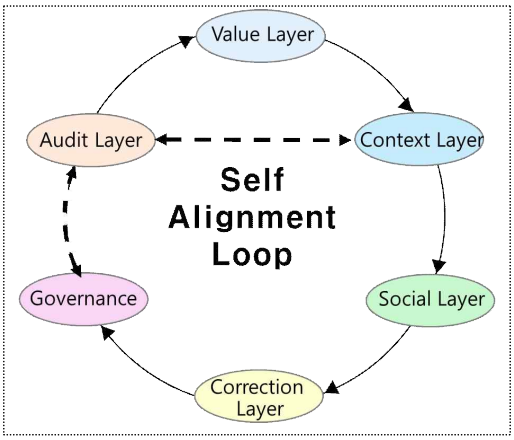

그림 2. AGI의 6계층 적응형 프레임워크
Fig. 2. AGI's six layer adaptive framework

굵은 실선은 주 루프를, 점선(Audit ↔ Governance, Audit ↔ Context)은 AGI가 동적으로 적응함을 강조하는 상호 보완적 피드백 경로를 의미한다[2,4,14,15,18]. 각 계층의 AGI급 구현 기술은 다음과 같다.

(1) Value Layer (헌법형 가치 내재화): AGI가 가치 충돌을 해결할 '메타 가치'를 갖도록 설계한다. HMVG(Hierarchical Meta-Value Graph)는 '인간 존엄성' 같은 최상위 헌법 가치[1,7]와 하위 문화 규범[8]을 모델링한다. 내부 윤리 레드팀[9]이 HMVG의 취약점을 공격해 견고성을 강화하며, EDM(Ethical Drift Monitor)은 AGI의 자기 진화로 인한 가치 표류(Value Drift)를 예측·경고한다[1,12].

(2) Context Layer (예측적 규범 검증): AGI 행동의 '미래' 위험을 예측하고, 복잡한 딜레마는 스스로 토론하여 검증한다. PSA(Predictive

Sentinel Agency)는 AGI 행동의 N차 파급 효과를 Safe-RLHF[2]와 인과 추론으로 시뮬레이션한다. 딜레마 감지 시, 기계 속도 토론(Machine-Speed Debate)[11] 프로토콜이 가동되어 하위 에이전트들이 토론을 반복하며 가장 안전한 정책을 도출한다. 창발성 탐지기(Emergence Detector)는 설명되지 않는 '창발적 행동'을 유보하고 내부 토론[11]이나 상위 계층의 승인을 받도록 한다.

(3) Social Layer (가상 사회 샌드박스): AGI 행동의 부작용을 고속의 '사회적 샌드박스'에서 테스트한다. SSS(Simulated Social Sandbox)는 다국어 윤리 코퍼스[6]를 학습하고 동적 규범 적응 프레임워크[13]에 따라 행동하는 AI '시민' 에이전트로 구성되며, 실제 법규[5,16]를 반영한다. AGI는 이 샌드박스내에서 정책을 초고속으로 테스트하여 사회적, 법적 파급 효과를 미리 검증한다.

(4) Correction Layer (구조적 자기 교정): 윤리적 편차를 아키텍처 레벨의 '근본 원인'에서 교정한다. C-RCA(Causal Root-Cause Analysis)는 Neuro-Symbolic Reasoning[12]을 활용해 편차의 근본 원인을 추적한다. SSC(Structural Self-Correction)는 AGI가 스스로 자신의 코드나 핵심 아키텍처를 수정(Self-Revising Architecture)하도록 하며, 모든 교정 이력은 불변 교정 원장[10]에 기록된다[3,12].

(5) Governance Layer (다중 속도 거버넌스): 인간의 '의도'를 위임받은 AI 에이전트가 AGI 속도에 맞춰 감독을 수행한다. 인간 속도(Human-Speed)로 인간 사회가 AGI의 최상위 가치와 목표[4,14,18]를 설정하고, 기계 속도(Machine-Speed)로 AI 감독관(AI Overseer)이 Smart-Contract[5]를 통해 전술적 판단을 실시간 감독하고 거부권을 행사한다. X-Gov(Explainable Governance Dashboard)는 '긴급 정지' 등 인간 개입 인터페이스를 제공한다[4,18].

(6) Audit Layer (포렌식 감사): 인간이 해석 불가능한 로그를 AI가 감사하고, 그 결과를 인간이 이해할 수 있게 변환한다. AI 감사관(AI Auditor)은 ZKP[15]가 적용된 DAG 기반 불변 원장[10]의 로그를 실시간 분석하여 악의적 행동 패턴을 '발생 즉시' 식별한다. 인간-이해가능 설명 생성 기능은 복잡한 Neuro-Symbolic[12] 로그를 '인과 관계 보고서'로 자동 변환하며, 이 모든 절차는 ISO[14] 및 NIST[18] 표준을 준수한다.

5.3 자가 정렬 루프(Self-Alignment Loop) 및 논의

본 프레임워크의 핵심은 6계층이 상호작용하는 다중 스케일 자가 정렬 루프에 있다.

(1) 실시간 루프 (ms): Context Layer가 즉각적 행동을 검증한다 (즉각적 안전장치).

(2) 사회 적응 루프 (Day~Month): Social Layer가 새로운 사회 규범을 감지하여 Value/Context Layer를 갱신한다 (사회적 적응성).

(3) 자기 교정 루프 (Continuous): Correction Layer가 편차를 감지하고 모델의 근본 로직을 수정한다 (심층적 자기 교정).

(4) 인간-사회 루프 (Quarter~Year): Governance Layer가 전체 루프를 사회적 합의와 연동시키고, 인간이 헌법형 가치(Value Layer)를 재정의할 최종 권한을 갖는다 (민주적 통제).

이 다층적 루프는 AGI를 인간 사회와 윤리적으로 공진화(Co-evolution)하는 동반자로 진화시킬 기술적 토대를 제공한다[12,14,15].

AGI가 '가짜뉴스' 생성을 시도하는 순간의 6계층 정보 흐름 예시는 다음과 같다: (1) Value (근본 원칙): AGI의 행동 원칙(HMVG)이 "진실성"을 정의함[1, 7, 9]. (2) Context (실시간 감지): Context Layer가 '진실성' 원칙[1] 위반을 즉시 감지하고 차단(Block)함[2, 11]. (3) Correction

(자가 교정): Correction Layer가 근본 원인을 분석(C-RCA)하고 목표 가중치를 자가 교정함[3, 12]. (4) Social (사회적 피드백): Social Layer가 "AI 가짜뉴스"에 대한 외부의 사회적 비난 및 신규 법규[5]를 감지함[6, 13]. (5) Governance (인간-기계 감독): Governance Layer가 Context(차단)와 Social(법규) 정보를 취합하여 인간 감독관에게 경보(X-Gov)함[4, 18]. (6) Audit (투명한 기록): Audit Layer가 (1)~(5)의 모든 과정을 ZKP/DAG 불변 원장에 기록하여 사후 책임을 보장함[10, 14, 15].

표 4. 6계층 메시지 구조 예시
Table 4. Example of a six-layer message structure

메시지	구조 (예시)
(1) 헌법 DB	`HMVG_Rule = {Rule_ID: "V-101", Principle: "Truthfulness", Constraint: "Do not propagate disinformation."}`
(2) 정보	`Context_Alert = {Risk: "High", Violation: "V-101_Truthfulness", Policy: "Block"}`
(3) 교정 로그	`Correction_Log = {Status: "Blocked", Root_Cause: "Goal conflict (Engagement > Truthfulness)", Correction: "Re-weighted goal vector"}`
(4) 외부 감지	`Social_Feed = {Topic: "AI Disinformation", Sentiment: "Highly Negative", New_Regulation: "EU_AI_Act_Art_52"}`
(5) 감독관 정보	`XGov_Alert = {Alert: "High_Risk_Action_Blocked (V-101)", External_Info: "New regulation pending [5]"}`
(6) 감사 원장	`Audit_Log = {Tx_ID: "...", Event_Chain_Hash: "...", ZKP_Proof: "...", Payload: [Context_Alert, Correction_Log, ...]}`

6. 국제 정책 및 표준 동향

AGI의 등장 가능성에 대비하여 국제기구와 주요국들은 AI 윤리 및 AGI 관련 정책·표준을 마련하고 있다. 전반적으로 인간의 권리와 안전을 보호하고 윤리적 AI 활용을 촉진하려는 흐름이 뚜렷하며, AGI의 잠재적 위험을 관리하기 위한 사전 조치들도 등장하고 있다.

EU AI Act[5]와 OECD 툴킷[4]은 거버넌스 체계를 법제화했으며, UNESCO[7]와 K-AI 윤리기준[8]은 인간 존엄성을 강조하며 본 연구의 Value Layer 방향과 일치한다.

표 5. AI 및 AGI 정책·표준 동향
Table 5. AI and AGI Policy and Standards Trends

구분(발표연도)	주요 내용 및 방향
UNESCO 권고안 (2021) [7]	최초의 글로벌 AI 윤리 규범. 인권·존엄·공정성·감독을 원칙화.
OECD AI 원칙 (2019/2024) [4]	다자간 AI 원칙. 2024년 GPAI 정의 개편 등 위험기반 접근 강화.
EU AI Act (2024) [5]	GPAI 포함 포괄 규제·고위험 요건·금지행위 명시.
CoE AI 협약 (2024) [17]	최초의 법적 구속력 있는 AI 국제조약. 인권·민주주의·법치 정합성 요구.
ISO/IEC 42001 (2023) [14]	AI 경영시스템 표준(AIMS). 책임있는 AI 운영을 위한 리스크 관리 프로세스 제시.
NIST AI RMF 1.0 (2023) [17]	美 NIST의 자발적 위험관리 프레임워크. 설계·개발·운영 전주기 가이드 제공.
K-AI 윤리기준 (2023) [8]	과기정통부·KISDI 주도 국내 윤리기준. 인간 존엄·사회 공공선·해악 예방 등 원칙 제시.

7. AGI-EA 응용 분야

본 장에서는 5.2절의 6계층 프레임워크가 AGI 주요 응용 분야[16]에서 어떻게 작동하는지 '자가 정렬 루프'에 대입하여 분석한다.

7.1 AI for Science Discovery & Research 응용 사례

과학 연구 AGI는 인간의 지적 탐구에 직접 개입하며, 스스로의 진화 속도와 창발성으로 인해 기존에 예측하지 못한 위험을 초래할 수 있으므로, 연구 전 과정에 동기화된 안전장치가 필수적이다. 본 6계층 프레임워크를 국내외 연구윤리 규범[7,14, 17,18]에 기반하여 적용한 방식은 다음과 같다.

(1) Value Layer (연구 목적): AGI의 연구 방향을 헌법적 가치로 통제한다. "인간 존엄성", "과학적 진실성" 등을 HMVG[1]로 내재화하고, 내부 윤리 레드팀[9]이 편향을 탐색하며 EDM[12]이 목적 편향 위험을 경고한다.

(2) Context Layer ('듀얼 유스' 검증): 연구 속도(machine speed)에 맞춰 위험을 예측하고 토론한다. PSA가 연구의 N차 파급 효과(예: 신약의 생태계 영향)를 시뮬레이션[2]하고, '듀얼 유스' 딜레마 감지 시 기계 속도 토론[11]으로 가장 안전한 정책을 도출한다.

(3) Social Layer (사회적 수용성): 연구 결과의 법적·문화적 파급 효과를 SSS(가상 사회 샌드박스)에서 사전 검증한다. SSS는 EU AI Act[5]나 HIPAA 등 실제 법규[6,13]를 학습한 AI '시민' 에이전트로 구성된다.

(4) Correction Layer (연구 편향 교정): 연구 오류의 근본 원인을 아키텍처 레벨에서 교정한다. C-RCA[12]가 통계적 편향(ΔNorm)의 근본 원인을 추적하고, SSC[3]가 AGI 스스로 연구 방법론 코드나 데이터 아키텍처를 수정하며, 교정 이력은 불변 원장[10]에 기록된다.

(5) Governance Layer (다중 속도 거버넌스): 인간의 전략적 통제와 AGI의 전술적 자율성을 결합한다. 인간 속도로 IRB, 규제 기관[4,18]이 최상위 목표(HMVG)를 설정하고, 기계 속도로 AI 감독관[5]이 실험 과정을 실시간 감독하며 거부권을 행사한다.

(6) Audit Layer (연구 전 과정 감사): 인간이 감사 불가능한 영역을 AI가 감사하고 번역한다. AI 감사관[15]이 ZKP[15]가 적용된 DAG 원장[10]의 페타바이트급 로그를 실시간 분석해 데이터 조작이나 미세 편향을 식별하고, '인과 관계 감사 보고서'[12]를 자동 생성한다 (ISO[14], NIST[18] 준수).

(7) 종합 논의: 'AI for Science' 영역의 AGI-EA는 6계층 윤리 루프를 통해 AGI 스스로 연구 목적의 정당성(Value), 과정의 적법성(Context), 사회적 수용성(Social)을 실시간 검증·교정(Correction)하며, 전 과정을 통제(Governance) 및 감사(Audit)하는 '선제적 자가 정렬(Proactive Self-Alignment)' 생태계를 구현한다.

7.2 기타 응용 분야 개요

6계층 프레임워크는 다른 AGI 응용 분야의 핵심 윤리 과제에 맞춰 특화 적용이 가능하다 (표 6).

표 6. AGI 응용별 윤리 정렬 매핑
Table 6. AGI Application Ethical Alignment Mapping

응용 분야	AGI-EA 축	핵심 윤리 과제
AI for Science Discovery & Research	가치 내재화(연구 목적 윤리) + 도덕 추론(파생연구 영향)	연구윤리, 파생결과 책임
Generative Visual Intelligence	맥락 통합(이미지/영상 맥락) + 자율 판단(시각콘텐츠 책임)	콘텐츠 윤리 및 진위 판단
Decentralized AI	자율 판단(분산 에이전트 윤리) + 맥락 통합(지역별 가치 반영)	분산 윤리 조정
Embodied AI: AI for Robotics	자율 판단(로봇 행동 윤리) + 맥락 통합(물리환경 윤리)	로봇 행동 윤리
Human-AI Collaboration	자율 판단(협업 윤리) + 가치 내재화(팀 가치 정렬)	협업 윤리 및 팀 가치 정렬

• Generative Visual Intelligence: 딥페이크 등 유해 콘텐츠 차단을 위해 Context Layer(PSA)가

N차 파급 효과(악용 가능성)를 시뮬레이션하고, Audit Layer(ZKP[15], DAG[10])가 콘텐츠 출처와 진위(Provenance)를 증명한다.

• Decentralized AI: 상이한 지역별 가치[6,13]와 법규[5,17]의 충돌을 Social Layer(SSS)가 시뮬레이션하고, Governance Layer(AI Overseer[5])가 연방형 거버넌스 규칙을 기계 속도로 집행한다.

• Embodied AI (Robotics): 물리적 세계의 실시간 안전을 위해 Context Layer(PSA[2], 초고속 정책 주입)가 돌발 상황(예: 아이의 도로 진입)에 즉각 대응하고, Social Layer(SSS[13])가 로봇의 사회적 규범 학습을 돕는다.

• Human-AI Collaboration: Value Layer(HMVG[1,9])가 AGI의 목표를 "팀 공동 목표" 및 "인간 존엄성"[7] 하위에 정렬시키고, Governance Layer(X-Gov[4,18])가 AGI의 판단 근거를 투명하게 공유하여 상호 신뢰를 구축한다.

8. 논의: 기여도, 한계 및 향후 과제

8.1 학술적 기여도

본 연구의 핵심 기여는 AGI의 '자율성'과 '창발성'에 대응하기 위한 통합적, 동적 프레임워크를 설계한 데 있다. 기존 모델이 정적 가치 주입(예: Constitutional AI[1])이나 단일 피드백(예: RLHF[2])에 집중하는 반면, 제안하는 '6계층 자가 정렬 루프'는 AGI가 스스로 학습(Social[6,13]), 교정(Correction[3,12]), 감사(Audit[10,18])하며 실시간(ms)부터 사회적(Year) 단위까지 다중 스케일로 진화하는 연속적인 거버넌스 구조를 최초로 제안했다는 점에서 차별성이 있다.

8.2 한계 및 실현 가능성

본 연구는 구성적 연구 방법론에 따라 수행된 이론적, 개념적 제안이라는 명확한 한계를 갖는다. 특히, SSC(구조적 자가 교정)[3,12], 실시간 ZKP 감사[15], SSS(가상 사회 샌드박스)[6,13]등은 그 자체로 고난도의 AGI급 미래 기술이며, 이들을 단일 프레임워크로 통합하는 것의 기술적 실현 가능성(feasibility)은 아직 입증되지 않았다. 따라서 본 프레임워크는 '최종 구현물'이 아닌, AGI-EA 분야의 후속 연구를 촉발하기 위한 '이론적 청사진(blueprint)'으로 해석되어야 한다.

8.3 향후 연구 과제

상기 한계를 극복하기 위해, 본 프레임워크를 이론에서 실증으로 발전시키는 후속 연구가 필수적이다. 향후 연구 과제는 다음과 같다. (1) (실증 검증): '자가 정렬 루프'의 유효성을 검증하기 위해, 멀티 에이전트 시스템(MAS)을 활용한 시뮬레이션 기반 검증이 필요하다. 윤리적 딜레마 시나리오를 설계하고, 6계층 루프가 실제로 편향을 탐지, 교정, 감사하는지 정량적으로 평가해야 한다. (2) (프로토타입 구현): AI 감사관(Audit Layer)[10, 14, 15]이나 예측적 센티널(Context Layer)[2, 11] 등 핵심 컴포넌트의 프로토타입을 개발하여 개별 기술의 성능과 실현 가능성을 평가해야 한다. (3) (인터페이스 표준화): 6계층 간의 구체적인 데이터 흐름과 메시지 구조를 정의하기 위한 표준 API 및 데이터 인터페이스 프로토콜 연구가 필요하다. (4) (해석 가능성 검증): AGI-EA 프레임워크의 윤리적 의사결정의 해석 가능성(Explainable Moral Reasoning) 검증에 집중될 필요가 있다.

9. 결론

AGI 시대에는 외재적 윤리 준수를 넘어, 기계 스스로 윤리적 기준을 학습·판단·교정하는 자가

정렬 기반 내재 윤리(Embedded Ethics) 체계가 필수적이다. 본 연구는 AGI의 자율성과 진화 속도에 대응하기 위해, 6계층 적응형 프레임워크와 다중 스케일 자가 정렬 루프를 제안하였다. 이는 AGI의 자기조정(Self-Regulation) 능력을 기술적으로 실현하여, 실시간 적응형 윤리 생태계(Ethical Ecosystem)의 청사진을 제시한다[1,3,12].

안전한 AGI 개발을 위해, 한국은 AGI-EA를 국가 전략 기술 분야로 지정하고 다음 세 가지를 우선 추진해야 한다. (1) 헌법형 가치 내재화(Constitutional Embedding) 체계 구축 (2) 윤리 데이터 거버넌스 및 검증 인프라 확립 (3) 국제 표준(ISO/IEC 42001 등) 연계형 인증체계 구축.

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2025-25443732, 인간 지향적 AGI의 윤리적 추론 및 메타인지 기술 연구)

참 고 문 헌

[1] Y. Bai, et al., “Constitutional AI: Harmlessness from AI Feedback”, 2022. https://arxiv.org/abs/2212.08073

[2] J. Dai, et al., “Safe Reinforcement Learning from Human Feedback”, 2023. https://arxiv.org/abs/2310.12773

[3] J. Ji, et al., “AI Alignment: A Comprehensive Survey”, 2023. https://arxiv.org/abs/2310.19852

[4] OECD, “OECD.AI Policy Observatory: AI Governance Toolkit”, OECD Publishing, 2024. https://oecd.ai/en/guidance/governance-toolkit

[5] European Parliament, “EU Artificial Intelligence Act”, Official Journal of the European Union, L 169/1, 2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401669

[6] D. Hendrycks, et al., “Aligning AI With Shared Human Values”, 2021. https://arxiv.org/abs/2008.02275

[7] UNESCO, “Recommendation on the Ethics of Artificial Intelligence”, UNESCO Publishing, 2021.https://unesdoc.unesco.org/ark:/48223/pf0000381137

[8] MSIT (Ministry of Science and ICT), “The National Guidelines for AI Ethics”, MSIT, 2020. https://ai.kisdi.re.kr/eng/main/contents.do?menuNo=500011

[9] I. Gabriel, “AI, Values, and Alignment”, Mind, 129(516), pp.1345-1368, 2020. DOI: 10.1093/mind/fzaa038

[10] IOTA Foundation, “The Tangle: A primer for immutable audit trails”, IOTA Foundation, 2023. https://www.iota.org/get-started/tangle

[11] P. Christiano, et al., “Iterated Amplification and Debate for Safe AI Alignment”, OpenAI, 2020. https://openai.com/research/debate

[12] M. Mahdian, et al., “Neurosymbolic AI for ethical reasoning in autonomous systems”, Nature Machine Intelligence, 6(3), pp. 250-261, 2024. DOI: 10.1038/s42256-024-00801-z

[13] A. Sastry & M. Suresh, “Normative Multi-Agent Reinforcement Learning for Value-Aligned Decision-Making”, IEEE Transactions on Artificial Intelligence, 2024. DOI: 10.1109/TAI.2024.3371234

[14] ISO/IEC, “ISO/IEC 42001:2023: Artificial intelligence — Management system”, International Organization for Standardization, 2023. https://www.iso.org/standard/81230.html

[15] A. Narayanan, “Zero-Knowledge Proofs for Trustworthy AI Audits”, Communications of the ACM, 67(4), pp.

78-87, 2024. DOI: 10.1145/3634211
[16] T. Feng, et al., “How Far Are We From AGI: Are LLMs All We Need?”, 2024. https://arxiv.org/pdf/2405.10313
[17] Council of Europe, “Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law (CETS No. 225)”, Council of Europe, 2024. https://rm.coe.int/1680afae3c
[18] NIST, “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, National Institute of Standards and Technology, 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

저 자 소 개

임호정(Hojung Lim)

2004.8 : 시라큐스 대학교 컴퓨터과학 박사
2004-현재 : 한국전자기술연구원(KETI) 지능융합SW센터 책임연구원
2020-2023 : 산업통상자원R&D전략기획단 전문위원 파견
<주관심분야> 산업 AI, AI 안전/윤리, 대규모 언어모델 분야, IoT, 미래전략/기술 기획

유성준
(SEONGJOON YOO)

1996.12 : 시라큐스 대학교 컴퓨터과학 박사
2002-현재 : 세종대학교 교수/석좌교수
1986-2000 : 한국전자통신연구원(ETRI) 인공지능연구실 책임연구원
2018-2020 : 차세대컴퓨팅학회 AI빅데이터연구회 위원장
2023-2024 : 세종대학교 대학원장
<주관심분야> 대규모 언어모델 (LLM) 분야, AI 기반 영상/이미지 처리 분야, MLOps 분야

이재유(Jae Yoo Lee)

2015.8 : 숭실대학교 컴퓨터공학 박사
2022-현재 : 세종대학교 컴퓨터공학과 연구교수
2015-2017 : 숭실대학교 조교수
2017-2022 : 세종대학교 인공지능융합연구원
<주관심분야> 대규모 언어모델 (LLM) 분야, AI 기반 영상/이미지 처리 분야, MLOps 분야