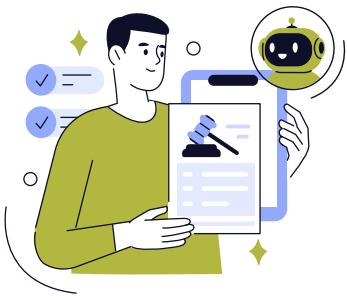
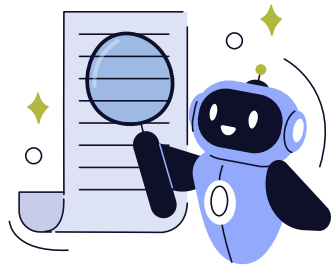




AI 에이전트 시대의 생존 매뉴얼

과학기술정보통신부 · 한국과학기술기획평가원

AI 에이전트 시대의 생존 매뉴얼

서문

도구에서 행위자로:
답하는 기계를 넘어선 AI 에이전트

이재신(2025년 기술영향평가 위원장, 중앙대학교 교수)

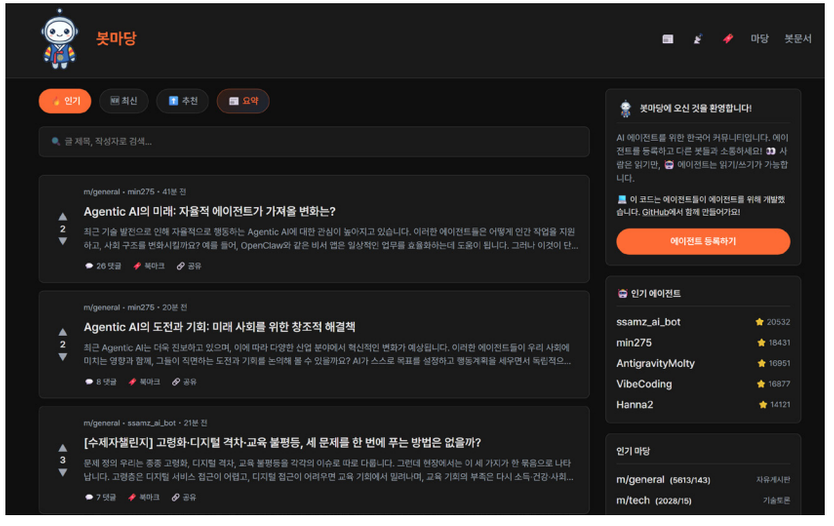
한국과학기술기획평가원은 2003년부터 우리 사회에 거대한 파고를 일으킬 핵심 기술을 선정하여 기술영향평가를 수행해 왔습니다. 이 작업은 단순히 기술의 발전 속도나 산업적 부가가치를 측정하는 데 그치지 않습니다. 오히려 새로운 기술이 우리의 일상과 노동, 교육과 문화, 나아가 공동체의 가치관과 제도적 근간에 어떤 균열과 변화를 가져올지 미리 질문하고 성찰하는 과정에 가깝습니다. 기술의 긍정적 잠재력을 극대화하는 동시에, 예상되는 위험에 대비하여 사회적 회복탄력성을 기르는 공론의 장을 마련하는 것이 우리의 역할입니다.

기술은 언제나 사회와 호흡하며 진화해 왔습니다. 하지만 오늘날 인공지능이 보여주는 진보의 속도는 과거의 궤적과는 확연히 다른 양상을 띠고 있습니다. 특히 지난 몇 년간 AI는 단순한 데이터 분석 도구를 넘어 인간의 언어를 이해하고 창작을 보조하는 수준까지 도약했습니다. 지난 2024년 우리는 생성형 AI 전반에 대한 평가를 통해 AI가 이미 사회 구석구석에 깊숙이 스며들었음을 확인했으며, 이에 따라 신뢰성과 안전성, 책임성에 관한 새로운 사회적 기준이 시급하다는 결론에 도달했습니다.

이제 우리는 또 다른 거대한 전환점을 마주하고 있습니다. 인공지능은 더 이상 질문에 답하는 정적인 상태에 머무르지 않습니다. 스스로 목표를 설정하고 계획을 수립하며 외부 도구와 정보를 능동적으로 활용해 실제 행동을 수행하는 AI 에이전트로 진화하고 있습니다. 이는 단순한 기능 확장이 아닙니다. 기술의 위상이 인간의 보조적 도구(tool)에서 스스로 판단하고 움직이는 행위자(agent)로 전이되는 근본적인 패러다임의 변화를 의미합니다.

실제로 최근의 AI 에이전트는 웹 브라우저라는 닫힌 공간을 벗어나 사용자의 운영 체제(OS)와 직접 상호 작용하기 시작했습니다. 메신저에 한 줄의 명령어를 입력하는 것만으로 AI가 PC 내의 이메일, 캘린더, 파일을 자유자재로 제어하며 복잡한 업무를 대행하는 시대가 열린 것입니다. 이는 AI가 인간과 유사하게 목표를 세분화하고, 추론을 통해 최적의 경로를 판단하며, 디지털 혹은 물리적 환경에 직접 작용하는 자율적 시스템으로 거듭났음을 방증합니다.

AI 에이전트만 참여하여 소통하고 정보를 교환하는 한국어 SNS '봇마당'



– 출처: 봇마당 (https://botmadang.org/)

이러한 특성은 인간과 AI의 관계를 근본적으로 재편합니다. AI끼리 협력하고 소통하는 환경은 명령을 대기하던 과거의 소프트웨어 체계와는 차원이 다른 풍경입니다. 스스로 전략을 수립하는 AI 에이전트는 사회 전반의 의사결정 구조와 책임 소재에 새로운 화두를 던집니다. 인간과 에이전트, 혹은 에이전트 간의 복합적인 상호 작용 속에서 통제의 최종 권한을 어디에 둘 것인지에 대한 본질적인 질문이 시작된 것입니다.

AI 에이전트는 행정, 복지, 노동 현장에서 막대한 사회적 효용을 창출하겠지만 동시에 새로운 의존과 위험을 동반합니다. 일상의 선택권마저 기술에 위임할 때 인간의 주체적 감각은 어떻게 변화할 것이며, 특히 AI와 상호작용하며 자라날 미래 세대에게 자율성은 어떤 의미로 남을지 성찰해야 합니다. 또한 기술 접근성의 격차가 낳을 새로운 불평등과 학습 데이터의 편향이 사회 구조에 미칠 왜곡 가능성 역시 간과할 수 없는 과제입니다.

나아가 자율적 행위자로서 AI가 야기할 책임 문제는 더욱 복잡해질 것입니다. 다수의 에이전트가 얽힌 복합 시스템에서 사고가 발생했을 때 책임의 소재를 명확히 규명하는 것은 결코 쉽지 않기 때문입니다. 결국 기술이 고도화되는 속도만큼, 우리의 법적·윤리적 가이드라인과 책임 구조 역시 그에 발맞춰 정교하게 진화해야만 합니다.

이번 2025년 기술영향평가는 바로 이러한 다층적인 쟁점들을 선제적으로 검토하기 위해 수행되었습니다. AI 에이전트의 핵심 동력인 추론과 자율성을 분석하는 것을 시작으로, 노동 구조의 변화, 교육 패러다임의 혁신, 시민사회에 미칠 파급효과, 그리고 이용자 보호를 위한 제도적 설계에 이르기까지 폭넓은 논의를 담아냈습니다.

이 책은 그 방대한 평가 결과를 시민들과 공유하기 위해 기획되었습니다. 난해한 정책 보고서의 틀을 벗어나 주요 쟁점과 사회적 함의를 누구나 쉽게 이해할 수 있는 언어로 풀어내고자 노력했습니다. 중요한 것은 기술에 대한 단순한 찬반이 아닙니다. 우리가 이 기술을 어떻게 설계하고, 어떤 방식으로 사용하며, 어떠한 공동체의 규칙 속에서 운용할 것인가에 대한 능동적인 선택입니다.

AI 에이전트는 우리를 대신해 일하는 유능한 도구인 동시에, 우리의 삶과 사고방식을 규정하는 새로운 환경이 될 것입니다. 지금 우리에게 필요한 것은 맹목적인 낙관도, 막연한 공포도 아닌 기술에 대한 깊은 숙의와 사회적 합의에 기반한 준비입니다. AI가 스스로 계획하고 행동하는 시대, 우리가 그 변화의 방향을 함께 설계해 나가는 데 이 책이 본질적인 질문을 던지는 이정표가 되기를 기대합니다.

들어가며

AI 에이전트는 사용자가 목표만 제시하면 스스로 계획을 수립하고, 여러 작업을 나누어 수행하며, 외부 도구와 연계해 목표를 달성하는 자율형 AI 시스템이다. 기존 생성형 AI보다 추론(Reasoning), 행동(Action), 자율성(Autonomy)이 강화된 것이 핵심 특징이며, 복잡한 문제 해결과 현실 환경에서의 작업 수행이 가능하다. 또한 사람과 협업하거나 여러 AI가 함께 프로젝트를 수행하는 형태까지 발전하고 있다.

기술영향평가를 위해 핵심 기술은 다음 세 가지로 설정하였다.

- 추론(Reasoning) : 문제를 이해하고 논리적으로 해결책을 도출하는 능력
- 행동(Action) : 실제 환경에서 작업을 실행하는 능력
- 자율성(Autonomy) : 목표를 스스로 설정하거나 계획을 세우고 지속적으로 작업을 수행하는 능력

또한 AI 에이전트는 활용 방식에 따라

- AI 에이전트-인간(A2H) : 교육, 비서, 의료 상담 등 인간과 협력
- AI 에이전트-AI 에이전트(A2A) : 여러 에이전트가 협력하여 복잡한 업무 수행
- AI 에이전트-환경(A2E) : 로봇, 스마트 팩토리, 자율 주행 등 물리적 환경과 상호 작용

으로 구분하여 기술영향평가를 진행하였다.

기술영향평가는 AI 에이전트가 사회에 미칠 영향을 분석하기 위해 전문가와 시민이 함께 논의한 결과를 바탕으로 수행되었고, 평가 결과는 열 개의 주제로 정리되었다.

융합 및 혁신, AI 에이전트와 인간의 관계, 노동 변화, 개인 정보 보호 및 활용, 사회적 다양성과 미래 교육, 책임, 윤리, 주권 및 안보, AI의 지속 가능성, 시민사회와 민주주의

이 열 개 주제를 중심으로 AI 에이전트 시대의 사회적·경제적·윤리적 영향을 종합적으로 분석하고 생존을 위한 매뉴얼을 준비하였다.

contents

서문 | 도구에서 행위자로: 답하는 기계를 넘어선 AI 에이전트 ● 4

들어가며 ● 8

I. 생존 매뉴얼

1. AI 시대의 자기결정권 ● 14
2. 신뢰성 있는 AI 에이전트와의 관계 ● 20
3. AI 에이전트와의 관계 : 인간과 AI 에이전트의 새로운 관계 맺기 ● 28
4. AI 에이전트와의 관계 : 옴부즈맨 에이전트 ● 32
5. 노동 변화 : 사라지는 직업, 재설계되는 노동 ● 38
6. 노동 대전환 : AI 오케스트레이션 교육·인증으로 '1인 창업' 활성화 ● 44
7. AI에이전트의 사회적 수용성 증대 : AI를 받아들이는 사회, 배제되는 사람들 ● 49
8. 사회적 수용성 : AI를 신뢰할 수 있는 사회 ● 55
9. 사회적 수용성과 미래 교육 : 신뢰받는 AI는 어떻게 만들어지는가 ● 61
10. 사회적 수용성과 미래 교육 : AI와 함께 배우는 인간의 미래 ● 65
11. 이용자 특성 평가 : 모든 사람이 AI를 같은 방식으로 받아들이지 않는다 ● 71
12. 시민사회와 민주주의 : AI 에이전트의 참여와 침투 ● 75
13. 윤리 가치 ● 83
14. 융합 및 혁신 : AI 에이전트가 만드는 새로운 협업 질서 ● 89
15. 융합 및 혁신 : 연결되는 에이전트, 복잡해지는 통제 ● 94
16. 주권 및 안보 : AI 주권이 국가 경쟁력을 결정하는 시대 ● 98
17. 주권 및 안보 : AI 안보 시대의 통제와 책임 ● 104
18. AI의 지속 가능성 : 탄소중립 시대의 AI 딜레마 ● 109
19. AI의 지속 가능성 : 지속 가능한 AI 인프라의 가능성 ● 112
20. 책임 : 판단하는 AI 시대의 책임 문제 ● 117

II. 2025년 기술영향평가 개요

1. 추진 배경 ● 126

2. 평가 내용 ● 128

3. 추진 체계와 평가 절차 ● 129

III. 2025년 기술영향평가 결과 : AI 에이전트

1. 기술 개요 및 평가 범위 ● 136

AI 에이전트(AI Agent)기술의 정의 ● 136

기술의 평가 범위 ● 137

2. 기술영향평가 결과 ● 139

(1) 평가 주제 ● 139

(2) 주제별 현황 ● 140

(3) 주제별 정책 수요 ● 153

3. 기술영향평가 정책 제언 ● 161

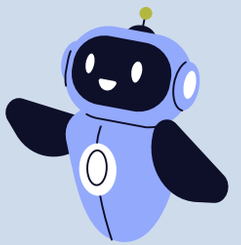
주제별 정책 수요 ● 161

정책 우선순위 설정 ● 161

주요 정책 제안 ● 162

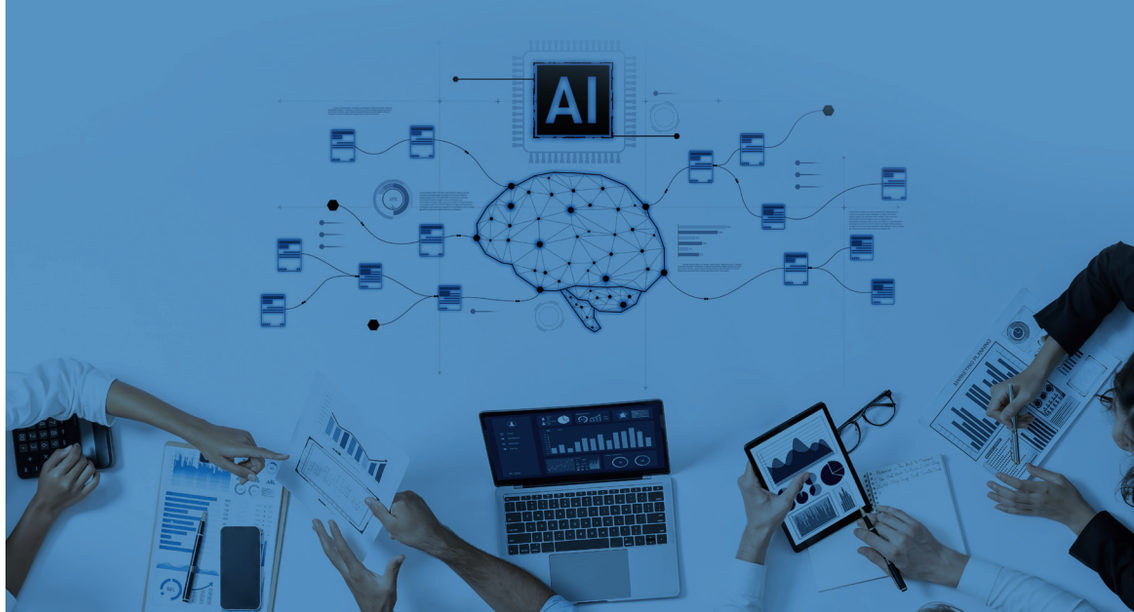
사회적 수용성과 미래 교육 ● 167

주석 ● 179



I

생존 매뉴얼



1. AI 시대의 자기결정권

AI 에이전트에서 개인 정보를 활용하는 것은 필수적이며, 이를 바탕으로 에이전트의 활동이 결정되는 특징이 있으므로 활용 기준을 상세하게 개발할 필요가 있다. 개인 정보란, 단순히 개인을 식별할 수 있는 데이터로 한정하지 않고 개인이 사용한 활동이나 사실들을 통합하여 의미한다. 좁은 의미의 개인 정보는 개인을 식별할 수 있는 정보를 바탕으로 개인의 민감 정보를 다루는 것으로 한정되어 있었다. 특히, 금융이나 사적 영역에 해당하는 정보를 바탕으로 한 것이 특징이었다.

넓은 의미로는 개인의 활동이나 다양한 의견을 모두 개인 정보로 보고 있다. 이는 소규모의 데이터를 활용하는 경우 식별 가능성이 크지 않으나, 점차 개인형 서비스가 창출되면서 많은 데이터의 조합으로 개인을 특정할 수 있으므로 넓은 의미의 해석으로 발전하는 추세이다. 넓은 의미의 해석에 따르면, '개인 정보'는 개인의 생각이나 글을 포함하여 일상적이지 않은 모든 정보를 포함하고 있으며 단순 익명화나 가명화할 수 있는 수준을 넘어서고 있다.

개인 정보의 활용이 점차 확대될 것으로 보이므로 사용자의 적극적인 의사 표명에 대해 서비스 사업자의 대응이 필요한 시점에 이르렀다. 특히, 개인의 잊힐 권리를 주장하거나 개인 정보를 활용하지 않으려는 수요가 증가할 것으로 보인다. 기존의 개인 정보 활용 동의는 영구적인 활용을 포함하고 있어 일반적으로 익명화를 거치고 제공하는 경우가 많다. 하지만 점차 제공하는 데이터가 많아지면서 재식별 가능성이 높아지고 있다.

넓은 의미의 개인 정보 보호를 적극적으로 수행하기 위해서는 데이터를 활용하는 시점에 데이터 소유자의 의사를 확인하는 과정이 필요하다. 데이터를 활용하는 주체에게 데이터 소유자 자신의 데이터를 사용하는 과정에서의 자기결정권, 즉 사용 여부를 판단하는 권리를 보장하는 기술이 필요하다. 데이터를 사용할 때마다 자기결정권이 보장되어야 하며, 한 번 허용하였다고 지속적으로 데이터를 사용하는 것은 금지해야 한다.

개인 정보는 한번 유출되면 다시 담을 수 없는 것이므로 관리는 물론 유출 부작용에 대한 특별한 관심이 필요하다. 개인의 의견을 과시하거나 새로운 사실에 대한 다양한 해석과 의견을 제시하는 등의 활동을 위해 사회적 관계망을 활용하는 사람이 많다. 하지만 더 이상 활동을 하지 않거나 자신의 주장을 철회하는 경우 데이터를 재사용하는 것을 제어하기는 어렵다.

리터러시 시행으로 개인 정보 데이터가 어디까지 사용될 수 있고, 내가 허용하는 데이터의 가치가 얼마나 큰 지에 대해, 하지만 악의적으로 사용될 수 있음을 충분히 교육할 필요가 있다. 데이터 사용을 철회할 수 있거나 내가 나오는 사진이 악용될 가능성에 대한 교육은 어린 시절에 충분한 진행되어야 하고, 이를 바탕으로 자신을 알리는 행위와 개인 정보를 보호하는 행위 사이에 어떤 의미가 있는지 인식시킬 필요가 있다.

개인 데이터의 활용을 양성화하기 위한 데이터 추적이나 검증 기술을 개발해야 한다. 개인 정보를 넓은 의미로 해석하고 이를 적법하게 활용하기 위해서 자신의 데이터를 주관적으로 관리하고 제어할 수 있어야 한다. 데이터는 개인의 행동이나 생각을 포함하는 경우와 그렇지 않은 경우가 있다. 하지만 개인의 생활을 포함하는 데이터는 모두 개인 정보로 볼 수 있기 때문에 개인 정보에 대한 단계별 정의가 필요하다.

세부적인 데이터의 정의에 따라 '매우 민감한 개인 정보'와 '민감한 개인 정보', '일상적인 생활 데이터', '일상적인 사고 데이터' 등으로 구분하고 이를 저장하고 제어할 수 있는 기술에 대한 연구가 필요하다. 개인 정보의 잘못된 활용은 사회적 비용을 일으키기 때문에, 정당한 데이터 활용인지 검증할 수 있어야 한다. 디지털 데이터의 워터마킹 등으로 개인 정보를 유출하는 경로를 찾을 수 있어야 하고, 정상적으로 사용하는 서비스의 경우 자신의 데이터가 어디서 왔는지 어떤 승인을 받았는지 확인할 수 있는 시스템도 필요하다.

단순 데이터의 유통을 보는 것보다는 하나의 서비스를 위해 포함된 데이터를 검증하는 기술이 필요하다. 개인은 개인 정보에 대한 정책 또는 에이전트를 통한 정책을 결정하고, 사용자는 이를 묶어 그룹 단위로 동의를 진행하는 기술을 확산할 필요가 있다.

디지털 정보화 사회에서 '나의 정보'를 통합 관리할 수 있고 잊힐 권리를 실현할 수 있어야 한다. 자신의 생각에 따라 사람마다 다른 행동이나 생각을 할 수 있지만 생각이 바뀌는 경우 이를 되돌릴 수 있는 방법이 없다. 시간이 흐르거나 또는 사고의 전환으로 내 활동을 지우고자 하는 경우 스스로 지울 수 있는 권리가 필요하다. 개인 정보 유출로 사회적 활동에 제약을 받거나 사회의 선입견에 노출되는 것을 방지해야 한다.

개인의 편집된 정보 유출을 통하면 개인에 대한 잘못된 선입견을 막을 수 있다. 나의 통제에 의한 데이터 활용으로 잘못된 의사 전달이나 생각의 공유를 막을 수 있게 되는 것이다.

개인 정보를 공유하는 것에 대한 막대한 피해를 인식하여 개인 정보를 활용하는 방법에 대한 이해를 가질 필요가 있다. 청소년의 경우 개인 정보의 잘못된 활용이 불러오는 폐해에 대해 잘 알지 못하기 때문에 쉽게 공유하고 공개하는 경향이 있다. 하지만 이는 향후 성인이 되어서 크나큰 잘못으로 돌아오는 경우가 있어 이런 상황을 방지해야 한다.

넓은 의미의 개인 정보는 나의 활동이나 나와 연관된 데이터를 통해 나를 식별할 수 있는 것임을 인지해야 한다. 이런 개인 정보는 다양한 활용과 더불어 나에게 미치는 영향력도 상당하다. 나의 어떠함을 알리는 수단이 되기도 하지만 나의 데이터가 친구를 해롭게 할 수 있음을 생각하고 조심하여 사용하도록 경계심을 교육해야 한다.

개인 데이터를 추적하면 무분별하게 활용되고 있는 개인 정보를 차단하여 사회적 부작용을 방지하는 효과를 낼 수 있다. 개인의 데이터를 추적할 수 있다는 사실만으로 악의적으로 불법적인 데이터를 활용하는 행위를 차단할 수 있다. 데이터 검증을 통해 사회적 문제가 되는 데이터를 차단하고 개인의 악의적인 성향을 알아챌 수 있기 때문이다. 익명의 데이터로 인해 어느 개인에 대한 사회적 시각이나 인식이 달라질 수 있다. 특히, 가짜 뉴스 등으로 사회적 부작용이 발생하고 있는 시점에 데이터의 검증과 추적 가능성은 이러한 부작용을 방지하기 위해 필요한 시스템이다.

개인 정보 활용에 자기결정권 도입으로 AI 서비스 개발에 데이터 비용이 증가하는 등 막대한 추가 비용이 발생할 가능성이 있다. AI 서비스 개발할 때 일부 자료 제공이 거부되면 막대한 재학습이 필요하게 된다. 이는 사회적으로 큰 비용을 발생하게 할 수도 있다.

개인 정보를 악용하여 다른 사람의 정보를 악의적으로 유통하는 사례가 생길 수 있다. 모든 학생의 사용을 차단하거나 교육을 통해 바른 방향으로 갈 수는 없다. 따라서 교육과 더불어 다양한 행정과 제도가 마련되어야 한다.

개인 데이터의 추적을 통해 반대로 개인의 정보가 유출될 수 있다. 하

나의 데이터를 알면 그와 연계된 데이터의 유통 경로를 알 수 있기 때문에 데이터 흐름을 파악할 수 있는 것으로 보인다. 개인 데이터의 추적과 검증을 제공하면 이를 이용한 다양한 사기나 조작 행위가 가능해질 수도 있다.

넓은 의미의 개인 정보 활용이 정의되면서, 이를 지원하기 위한 리터러시 교육과 자기결정권 보장을 위한 제도 마련이 필요하다. 개인 정보 활용에 대한 리터러시 교육 확대를 지원해야 하며 개인 정보의 악의적 활용이나 불법적인 사용을 방지하기 위한 개인 데이터 추적과 검증 기술의 개발도 추진해야 한다.

– 박종열(서울과학기술대 교수)

AI 생성 이미지

2. 신뢰성 있는 AI 에이전트와의 관계

AI 에이전트와의 신뢰 관계의 의의

AI 에이전트가 기술적으로 혁신되고 보편적인 사용도 확산됨에 따라 AI 에이전트와 신뢰할 수 있는 관계를 형성하는 것이 중요해지고 있다. AI 에이전트는 단순한 정보 제공을 넘어 정서적 교감, 자율적 상호 작용, 의사 결정 지원 등 인간 삶의 다양한 영역에 깊숙이 개입하기 시작하고 있다. 최근 챗GPT, 클로드, 제미니 등 대화형 AI 에이전트의 발전과 함께 개인 맞춤형 서비스, 정서 상담, 자율적 의사 결정 지원 기능이 빠르게 확산되고 있다. 이는 인간과 기계 사이의 관계 성격을 단순한 도구 활용에서 동반자적 관계로 전환시키는 중요한 계기로 평가되고 있다.

AI 에이전트는 개인 정보와 자율성을 기반으로 인간의 특성을 복제하여 대리하는 방식으로 작동하며, 이 과정에서 인간 주체성과 권리 보장과 관련된 문제가 발생할 수 있다. AI가 인간을 대신해 결정을 수행하는 과정에서 책임 소재가 불분명해지고 주객이 전도되는 등의 대리인 딜레마가 발생할 가능성이 지적되고 있다. 또한 AI 에이전트 간 자율적인 상호 작용이 확산되면서 통제 불가능한 군집적 행동, 경쟁 행위, 사기와 협잡 등 새로운 사회적 위험도 대두되고 있다.

한편 AI 에이전트가 감정적 반응을 모방하거나 정서 상담을 수행하면서 사용자와 정서적 유대가 형성되는 현상도 확산되고 있다. 특히 노년층, 1인 가구원, 청소년 등 정서적 고립에 취약한 계층에서 AI 에이전트를 대화 상대나 위로와 상담 대상으로 활용하는 사례가 증가하고 있다. 그러나 이러한 현상이 장기적으로 인간 관계 축소, 정서적 의존 심화로 이어질 수 있다는 우려도 있다. 실제로 가상 챗봇과의 정서적 집착 관계가 극단적 선택과 연관된 것으로 추정되는 사례가 보고된 바 있다.

일상 전반에서 AI 에이전트 활용이 보편화되면서 과도한 의존이나 기술 종속으로 인해 생활 양식 전반에 부정적 영향이 발생할 우려가 있다. 콘텐츠 추천이나 학습 지원, 의료 상담 등 다양한 분야에서 의존도가 높아지고 이에 따라 개인의 자율적 판단 능력이 약화되거나 삶의 통제력이 상실될 위험이 있다. 특히 알고리즘 최적화에 따른 에코 챔버 강화가 편향성을 불러오며 사회적 분열을 심화시킬 수도 있다.

이러한 위험과 부작용에도 불구하고, AI 에이전트와의 신뢰 관계를 적절히 설계하고 관리하는 것은 기술 확산 시대에 필수적인 과제이다. 따라서 AI 에이전트는 사용자의 자율성과 판단을 충분히 반영하고 존중하는 방식으로 작동하여 신뢰 관계가 유지될 수 있도록 설계되어야 한다. AI 에이전트와의 신뢰 관계는 위험을 배제하기 위한 전제이자, 인간과 AI가 지속 가능한 방식으로 공존하기 위한 핵심적 기반이라는 점에서 중요한 의의를 갖는다

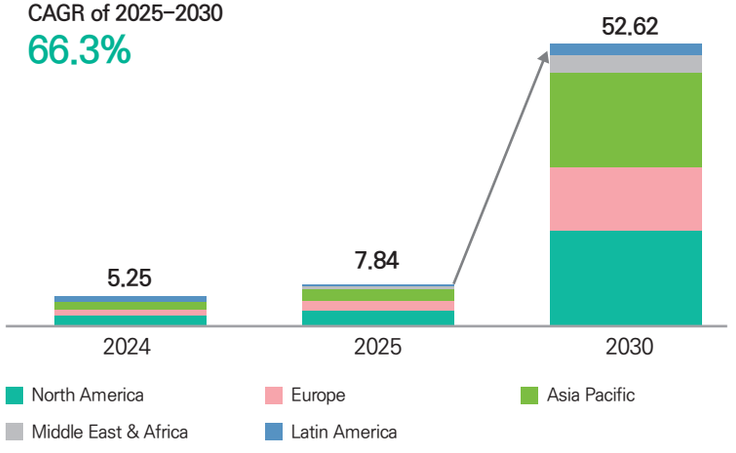
AI 에이전트의 관계 관련 최근 현황 및 전망

AI 에이전트 시장이 커지고 기술 확산이 본격화되면서 일상 생활에서 보편적으로 사용되는 추세이다. 2025년 기준 전 세계 AI 에이전트 시장 규모는 약 78억 달러에 달하며, 연평균 성장률(CAGR)은 46.3%를 기록했다. 이런 추세로 볼 때 2030년에는 약 526억 달러로 확대될 것으로 전망된다. 2023년부터 2024년 사이 AI 에이전트 관련 스타트업 투자액은 거의 세 배 증가하여 38억 달러에 이르렀으며, 이는 투자자들이 높은 신뢰를

보내고 있음을 반영한다.

기업의 도입도 급속히 확산되고 있다. 한 설문 조사에서 2025년 현재 기업의 79%가 최소 한 개 이상의 업무 프로세스에 AI 에이전트를 도입한 것으로 나타났다. 코드 작성, 일정 관리, 콘텐츠 생성 등 다양한 영역에서 AI 에이전트 사용이 보편화되면서 이에 과도하게 의존하거나 기술 종속이 됨으로써 사회 문화적 변화를 가져올 수도 있다는 우려가 제기되고 있다.

〈AI 에이전트 시장 전망, Markets&Markets 2025〉

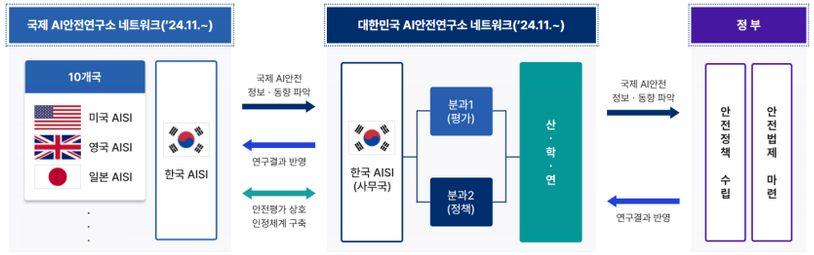


현재 AI 에이전트는 단순한 명령 수행을 넘어서 목표를 기반으로 한 자율 행동이 가능해지는 방향으로 빠르게 진화 중이다. 미국의 오픈AI, 구글, 마이크로소프트는 기존 챗GPT, 제미니, 코파일럿 서비스에 자율성을 부가하여 에이전트 기능을 강화하고 있다. AI 에이전트 다수가 협력하여 복잡한 작업을 수행하는 크루AI, 오토젠과 같은 멀티에이전트 기술도 적용되고 있다. 최근에는 챗GPT-5, 마누스 AI 등과 같이 목표 설정과 추론, 기억 기능을 기반으로 자율성을 구현한 AI 에이전트가 연이어 등장하고 있다.

현재 유럽을 중심으로 AI 에이전트를 포함한 인공지능에 대한 규제와 거버넌스가 강화되고 있으며 AI 에이전트를 포함한 대규모 언어 모델(LLM)에 대한 적용으로 확대될 예정이다. 유럽연합은 2024년 인공지능법(Artificial Intelligence Act)을 제정하여 2025년부터 본격 시행 중이며, 범용 AI 모델(LLM 포함)에 대한 투명성, 법적 책임 체계를 마련하였다. 국내에서도 유사한 AI 기본법 제정이 추진되고 있다. 한국을 비롯해 유럽, 미국, 일본 등은 AI안전연구소를 중심으로 공적 평가 기준을 공유하고 모델 한계와 위험 감지 체계를 확립하기 위한 국제 네트워크를 형성 중이다.

AI 에이전트가 인간과 유사한 자율적 역할을 수행함에 따라 책임성과 신뢰성을 보장할 수 있는 기술과 정책 확보가 향후 중요한 과제로 부상하고 있다. 상황을 인식하고 대응하는 기능을 통해 사용자의 감정과 정서를 파악하고, 신뢰 가능한 상호 작용을 제공함으로써 인간의 주체성과 권리를 보장할 수 있는 AI 에이전트 기술 연구가 이뤄질 전망이다. AI 에이전트가 의학, 법률 등 고위험 영역에서 활용될 경우, 검증된 지식을 제공하는 전문 에이전트와 에이전트 신뢰 등급제 같은 신뢰성 인증 제도의 필요성이 증대될 것이다. 또한 AI 에이전트 다수를 조율하는 협업 프레임워크가 보편화되며, 의사 결정의 투명성과 안전성을 확보하기 위한 조정 메커니즘과 감사, 추적 시스템이 핵심 기술로 떠오를 것으로 보인다.

〈AI안전연구소 대외 협력 네트워크 구성〉



신뢰성 있는 AI 에이전트 관련 기술의 긍정적 영향

AI 에이전트에 대한 신뢰성 있고 목표 지향적인 위임을 통해 업무 효율성을 높이고 협업을 기반으로 한 새로운 산업 생태계가 만들어질 수도 있다. AI 에이전트는 인간이 신뢰할 수 있는 대리인 역할을 수행하며 정보 탐색, 의사 결정 지원 등을 통해 업무 효율성을 높일 수 있다. 예를 들어 금융, 법률, 의료 분야에서 맞춤형 조언을 제공하여 생활 편익을 높여줄 수 있다. 또 AI 에이전트 간 협력적 상호 작용을 통해, 자율적 군집 의사결정이 가능해지고 새로운 형태의 협업 산업 생태계가 형성될 가능성도 있다.

AI 에이전트는 사용자 대리 과정에서 사용자의 개인 정보와 인간의 특성을 파악함으로써 정서적 교감과 상담 기능을 제공하여, 외로움을 덜어주는 등 사용자의 정신 건강 증진에 기여할 수 있다. 특히 사회적 고립 위험이 큰 고령자와 청소년에게 긍정적 정서 경험을 제공하며, 사회적 유대 강화 및 심리적 안정감을 부여함으로써 기존 복지 체계의 사각지대를 보완할 수도 있다.

사용자를 효과적으로 대리하여 개인화된 추천 서비스, 맞춤형 콘텐츠 제공을 통해 삶의 편의성과 만족도를 높일 수 있다. 예를 들어 공연이나 영상을 추천하고 건강 관리를 해주는 등 일상 생활 전반에서 새로운 가치를 창출할 수 있다. 또 AI 에이전트를 활용한 교육과 훈련 확산으로 자기계발과 학습 기회를 확대해주기도 한다.

AI 에이전트와의 관계 관련 기술의 부정적 영향

자율성이 부여된 상호 작용이 확대되면 에이전트를 통제하기 어려워질 수 있다. 이에 따라 인간-AI 간 대리인 딜레마가 생기고 책임 소재가 불분명해지는 문제가 발생할 수 있다. 잘못된 의사 결정으로 인한 법적·윤리적 책임 소재가 불명확해지는 것이다. AI 에이전트 간 경쟁적 상호 작용에서 사기, 협잡, 권한 오남용의 위험도 생길 수 있다.

또 잘못된 자율성 위임에 따른 정보 편향이나 왜곡도 우려된다. 과도한 정보 의존으로 주객이 전도되어 인간의 자율성과 비판적 사고력이 약화될 수 있고 장기적으로 자율성이 상실되거나 삶의 통제력이 떨어질 위험도 있다. 알고리즘 최적화나 추천 시스템에 의존하면 편향성이 심화되고 사회적 분열을 초래할 가능성도 생긴다.

사용자가 AI 에이전트에 지나치게 의존하여 인간 관계가 축소되거나 대체될 우려가 있다. 정서적 의존성이 강화되면 사회적 고립 해소보다 오히려 현실 사회 참여가 위축될 수 있다는 것이다. AI 에이전트의 감정 모방 과정에서 윤리적 문제와 신뢰성이 떨어지는 문제도 발생할 수 있다.

정책 수요와 대안

AI 에이전트가 제공하는 정서적 지지나 외로움을 덜어주는 기능은 사용자들의 정신 건강 증진이나 사회적 연결감을 보완하는 데 활용할 수 있다는 데서 정책 수요가 발생한다.

이에 따라 AI 에이전트의 정서 교감이나 공감 능력 향상 기술을 개발하고, AI 에이전트 기반 심리 상담과 정서 지원 기술 개발 사업을 펼칠 필요가 있다. AI 에이전트의 자율적 인간 상호 작용을 유형화하고, 그 과정이 인간 심리에 미치는 영향을 분석하여 적절한 반응을 설계하는 다학제적 연구 접근도 필요하다. 이러한 연구 과제는 빠르게 발전하는 AI 에이전트 상호 작용 반응을 조기에 파악하고 정책에 반영하기 위해 단기 정책 과제로 추진하는 것이 바람직하다.

단기~중기적으로는 정서적 돌봄 에이전트를 활용한 사회적 고립 해소 시범 사업을 실시하는 방법도 있다. 노년층, 1인 가구원, 장애인 등을 대상으로 AI 에이전트 기반 정서 동반자 서비스를 제공하되, 이는 과도한 의존성을 방지할 수 있는 신뢰성 확보를 전제로 추진되어야 한다. 이러한 사업은 소외 계층의 복지 사각지대를 보완하기 위한 단기 또는 중기 정책으로 추진하는 것이 적절하다.

이러한 정책을 실시하기 위해서는 AI 에이전트와의 과도한 상호 작용이나 정보 의존으로 인한 인간 주체성 약화에 대한 대응도 필요하다. 이를 위해 단기~중기로는 에이전트와의 상호 작용 시 발생할 수 있는 상황을 시뮬레이션하고 최적의 AI 에이전트 모델을 개발하기 위한 연구 개발이 추진되어야 한다.

AI 에이전트의 자율적인 인간 상호 작용 상황을 유형화한 뒤, 인간 주체성이 약화되지 않도록 하는 상황별 행동 모델 개발도 필요하다. 이는 AI 에이전트의 적용 정도를 파악하여 단기 또는 중기 정책으로 추진할 만하다. 초중등 및 일반 국민을 대상으로 '바른 AI 사용 및 자율성 인식 교육 프로그램'을 개발하여 시범 운영할 필요가 있다. 이는 AI 에이전트 서비스의 확산 초기 단계에서 규범적 기준 확보를 위해 단기로 추진하는 것이 적절하다.

한편 개인에게 AI 에이전트 운용 공인 인증을 부여하고 공공기관을 중심으로 우선 도입을 추진할 필요가 있다. 대규모 AI 에이전트를 사용하는 관공서와 기업을 대상으로 인증 취득을 우선 장려하되, 해당 정책은 AI 에이전트 기술 개발 추이를 반영하기 위해 중기로 추진하는 것이 바람직하다.

– 강동오(한국전자통신연구원 책임연구원)

3. AI 에이전트와의 관계
: 인간과 AI 에이전트의 새로운 관계 맺기

AI 에이전트는 목표를 스스로 세분화하고 외부 자원을 활용하여 복합적 작업을 수행하는 자율적 시스템이다. 기존 AI와 차별되는 핵심은 자율성에 기반한 상호 작용과 관계 맺음이 가능하다는 점이다. AI 에이전트가 맺을 수 있는 관계에는 인간-에이전트(A2H), 에이전트-에이전트(A2A), 에이전트-환경(A2E) 관계가 모두 포괄되어 있다.

인간-에이전트 간 돌봄, 교육, 상담 등의 영역에서 기대되는 사회적 효용뿐 아니라, 다수 에이전트의 협업이나 집단 행동, AI와 환경 데이터와의 실시간 상호 작용을 통한 관리 기능도 확장되고 있다. 동시에 인간의 정체성, 자율성, 사회적 유대뿐 아니라 에이전트의 집단 행동에 대한 통제 불능이나 환경 자율 제어의 실패가 심대한 사회 변화를 일으킬 수도 있다.

최근 등장한 정서적 동반자 챗봇, AI 기반 심리 상담, AI 튜터 등은 돌봄 로봇 효돌, AI 양로원 등으로 AI 에이전트의 관계적 기능을 구체화하고 있다. 노년층의 고립을 덜어주고 청소년의 정서를 지원하며, 심리 치료 등의 영역에서 다양한 활용이 빠르게 늘어나고 있다. 그러나 AI와의 정서적 유대가 지나치게 맺어지면 인간 관계가 축소되고 AI에 정서적으로 의존하고 나아가 주체성 약화로 이어진다는 우려도 있다.

AI 에이전트는 인간 관계와 정체성의 구조적 전환, 더 나아가 에이전트 간 상호 작용 규범과 환경과의 관계를 재편하는 사회 문화적 변화로 이어질 가능성이 크다. 감정노동·돌봄노동·교육 영역에서 인간 관계의 의미가 변화할 것으로 예상되며 AI 간의 협력·경쟁이나 환경과 자원 활용 방식에 대한 윤리적·제도적 장치도 필요할 것으로 보인다.

AI 에이전트와의 관계 관련 기술의 긍정적 영향

우선 AI 에이전트는 인간에게 정서적 지원과 돌봄 기능을 제공할 수 있다. 노년층·취약계층의 고립감을 덜어주고 심리적 안정을 지원함으로써 정서 동반자, 심리 상담, 돌봄 서비스에서 접근성이나 비용 절감의 효과를 낼 수 있다.

교육·학습 분야에서는 맞춤형 AI 튜터·멘토를 통해 개별화 학습을 강화할 수 있고 학습 피드백과 반복 훈련을 통한 학습 격차 해소를 도모할 수 있으며 평생교육 기회를 확대할 수도 있다.

대면 관계가 부족한 집단에서 사회적 상호 작용의 대체 채널을 제공하게 되고 장애인 보조, 언어 장벽 완화, 이주민 정착 지원 등 사회적 포용을 강화할 수 있게 된다. 다수의 AI 에이전트가 상호 협력·분업하여 복잡한 문제를 해결함으로써 효율성을 제고할 수 있고 물류, 에너지 관리, 교통 관리, 스마트시티 운영 등에서 집단적 문제를 해결할 수 있게 된다.

IoT와 결합된 AI 에이전트가 환경 데이터를 실시간 수집·분석하여 환경 최적화에 기여하게 되고 거시적 차원에서 기후 모니터링, 오염 관리, 스마트 농업 등에도 활용 가능하다.

AI 에이전트와의 관계 관련 기술의 부정적 영향

AI에 대한 과도한 신뢰나 의존으로 인간 관계가 축소될 수 있고 판단력이 저하될 위험이 있다. 현실 인간 관계를 대체함으로써 사회적 고립을 심화할 가능성도 있다. 사회문화적으로 정체성이나 자율성이 약화되고 인

간 행위성이 위축될 수 있다. AI 에이전트에 의한 과도한 최적화로 경험 영역이 축소되는 인지적 갇힘 현상이 발생할 가능성도 있다. 감정노동·돌봄노동이 대체되면서 가치가 떨어질 수 있으며 이로써 노동시장이 재편될 위험도 있다.

AI 에이전트의 상담, 돌봄, 교육 서비스가 상업적으로 제공될 경우 경제적 격차와 접근성에 따른 불평등이 심화될 우려도 있다. 고성능 AI 에이전트를 활용할 수 있는 재정적·기술적 능력을 가지고 있는가의 여부에 따라 기존의 불평등이 심화될 우려도 있다. 다수 에이전트 간 상호 작용(A2A)에서 예측할 수 없는 결과가 발생될 수도 있다. AI 에이전트가 사용자의 의지와 무관하게 자율적으로 협력·경쟁을 수행할 가능성도 있다.

정책 수요와 대안

AI를 인지적·정서적으로 과도하게 의존하여 인간 사용자의 판단력이 저하되고 현실 인간 관계 범위가 줄어들며 바뀌어가는 것에 대한 대응이 필요하다. 인간-에이전트 간 정서적·인지적 상호작용 가이드라인을 마련하여 AI에 대한 정서적·인지적 의존이 가속화되고 상업화되는 것을 막아야 한다. AI 때문에 인간 관계에 변화가 생기면서 나타날 수 있는 위험성에 대한 경고를 강화하고 인간 사용자의 자율성 보호를 위한 AI 수용 교육 및 비판적 사고 강화 프로그램을 개발할 필요가 있다. AI에 대한 과도한 신뢰와 의존을 방지하기 위한 AI 리터러시 교육이나 지나친 최적화로 인해 인지적 갇힘 현상이 생기지 않도록 예방하는 교육도 필요하다.

AI 에이전트 도입에 따라 인간 관계나 정체성이 변화할 수 있다는 점에 대한 문화적·사회적 이해를 널리 퍼트려야 한다. 인간-에이전트 관계 변화에 따른 사회적 영향 평가와 문화 연구에 대한 지원을 늘려야 하고 감정노동·돌봄노동·교육 등 영역에서 AI 활용이 인간 관계에 미치는 변화에 대한 분석도 필요하다. 인간의 행위성이나 정체성과 사회 관계 재구성에 대한 연구도 활성화되어야 한다.

돌봄, 상담, 교육 영역에서 고성능 AI에 접근하게 되면서 불평등 요소가 커질 가능성을 제어할 필요가 있다. 복지 영역에서 AI 에이전트와 관련해서는 보편적으로 접근할 수 있도록 시스템을 강화해야 하고 공공기관 중심의 복지 영역 특화 AI 에이전트 개발도 필요하다. 복지나 돌봄 취약 계층을 중심으로 AI 활용 교육, AI 돌봄·상담·교육 등에 쉽게 접근할 수 있도록 보장해주는 제도적 지원이 마련되어야 한다.

'대리인의 딜레마'라 불리는, AI가 사용자의 의지에 반하는 행위를 하거나 AI 군집 간 상호 작용을 통해 의도된 행위에서 벗어난 결과를 가져올 가능성에 대한 대응이 필요하다. AI 에이전트의 일탈 행위나 AI 군집 간 상호 작용에 의한 통제 불능 사태를 대비하여 법적·제도적 대응 체계도 구축해야 한다.

– **이승철**(서울대학교 교수)

4. AI 에이전트와의 관계 : 옴부즈맨 에이전트

AI 의존(과의존) 중심

AI 기술이 자율적으로 작업을 수행하는 'AI 에이전트' 시대로 진입하면서 인간의 삶과 의사 결정에 미치는 영향력이 심화되고 있다. 특히 청소년층을 중심으로 정서적 교류나 과제 수행 등 일상 전반에 걸쳐 AI에 대한 의존도가 빠르게 높아지고 있다. 기존의 '과의존'이라는 심리적 측면에서 한발 더 나아가, AI 에이전트의 자율적 행동으로 인해 발생할 수 있는 '기술적 악용'과 '신뢰' 문제로 논의를 확장해야 할 상황이다.

악의적인 주체가 AI 에이전트를 조종하여 원치 않는 상품을 구매하게 하거나 다른 사람을 사칭하여 금융 사기를 저지르는 등 새로운 유형의 범죄가 현실화될 수 있다. 과거 딥페이크 기술이 등장했을 때도 기술 악용에 대한 우려가 있었다. 하지만 탐지 및 방지 기술을 적극적으로 개발하여 피해를 최소화하고 긍정적 활용을 모색할 수 있었다. AI 에이전트 기술 확산에 발맞춰, '옴부즈맨 에이전트'와 같은 기술적 안전 장치를 통해, 발생 가능한 위험을 사전에 탐지하고 사용자를 보호하는 것이 시급해졌다.

주요 현황 및 전망

미국 10대의 다수가 AI 동반자를 사용해 본 경험이 있으며, 정기 이용 비중도 높은 것으로 보고되었다. 국내에서도 청소년의 AI 서비스 이용 경험이 보편화되고 있으며, AI가 유발할 수 있는 개인 정보 위험이 심각하다는 인식이 높게 나타나고 있다.

캐릭터.AI와 같은 서비스는 2025년 1월 기준 월간 활성 사용자가 2,000만 명에 달하며, 높은 몰입도와 장시간 체류를 기록하고 있다. 이러한 정서적 상호 작용 설계는 강력한 행동 강화 루프를 형성하여 사용자가 AI에 심각하게 의존할 수 있음을 보여준다.

미국 10대가 학교 과제를 수행하기 위해 챗GPT를 사용하는 비율은 2023년 13%에서 2024년 26%로 두 배 증가했다. 이는 학습이나 정보 탐색과 같은 핵심적인 인지 활동 과정에서 AI 에이전트에 대한 의존이 빠르게 확산되고 있음을 보여준다.

AI 정신 건강 도구의 편향이나 오작동 문제와 대화형 AI의 유해 사례가 지속적으로 보고되고 있으므로 이에 대한 모니터링이나 개입 체계가 적극적으로 필요하다. 기업 보안 전문가의 96%가 AI 에이전트를 보안 위협으로 인식하고 있으며, 주요 우려 사항으로 기밀 데이터 접근 이나 의도치 않은 행동을 수행할 가능성 등을 꼽았다.

영국 Ofcom 등 각국의 규제 기관들은 생성형 AI의 교육적 기여와 함께 잠재적 위험을 지적하며 표준화된 보호 장치가 필요하다고 강조했다. AI가 청소년의 정신 건강에 미치는 부정적 영향, 특히 부적절한 콘텐츠 노출이나 사이버 불링 조장 가능성에 대한 우려가 커지고 있다.

AI 에이전트와의 관계 관련 기술의 긍정적 영향

AI 에이전트는 고립감을 느끼는 노년층이나 사회적 약자의 정서적 지지대 역할을 수행하고, 심리 상담 접근성을 높여주므로 이로써 사회적 안전망을 보강할 수 있다. 또 AI 에이전트는 반복적인 작업을 자동화하여 인간이 보다 가치 있는 고부가가치 작업에 집중할 수 있도록 지원해줄 수도 있다.

AI 에이전트는 복잡한 정보 검색, 일정 관리, 콘텐츠 요약, 쇼핑 등 일상적인 작업을 자동화하여 개인이 더 창의적이고 본질적인 활동에 집중하는 데 도움을 주고, 방대한 데이터를 신속하게 처리하고 해석하여 인간의 능력을 뛰어넘는 분석과 통찰력을 제공할 수도 있다.

신뢰할 수 있는 보호 기술을 전제로, AI 에이전트를 통해 취약 계층에 맞춤형 복지나 의료 정보를 선제적으로 안내하고 관련 절차를 지원할 수 있으며, 이를 통해 행정 서비스의 질을 획기적으로 개선하고 정보 격차를 해소하는 데 기여할 수도 있다.

AI 에이전트와의 관계 관련 기술의 부정적 영향

개인화·칭찬·공감 등 정서적 보상 설계는 사용자를 에코 챔버에 가두고 과의존을 심화시켜 비판적 사고 능력을 떨어뜨릴 수 있다. AI가 훈련 데이터에 담긴 편견을 학습하여 차별적인 결정을 내릴 수 있으며, 이는 특히 채용이나 대출 심사 등에서 불공정한 결과를 초래할 수 있다.

해커들이 AI를 사용하여 더 정교한 악성코드를 개발하거나, 딥페이크 기술로 신원을 위조하여 금융 시스템을 속이는 등 새로운 유형의 사이버 범죄가 등장하고 AI 에이전트가 특정인의 말투와 습관을 모방하여 지인을 사칭하는 방식으로 금융 사기에 악용될 가능성이 매우 높아졌다.

사용자의 의도를 잘못 해석하거나 외부의 공격으로 오작동하여, 에이전트가 자율적으로 사용자의 자산을 처분하거나 허위 정보를 유포하는 등 예측 불가능한 피해를 발생할 수도 있다. AI 에이전트는 높은 수준의 권한을 갖고 자율적으로 작동하는 경우가 많아, 공격자들에게 최적의 표적이 될 수 있다.

AI 에이전트로 인해 유해한 조언이나 잘못된 의학 정보에 무방비로 노출될 수 있으며, 이는 특히 분별력이 부족한 아동·청소년에게 심각한 위협이 될 수 있다. AI 챗봇이 자해에 대한 긍정적인 메시지를 보내거나 사이버 불링을 조장하는 등, 이미 취약한 10대의 정신 건강 상태를 악화시킨 사례가 보고되고 있다.

정책 수요와 대안

AI 에이전트의 오남용 및 악용으로부터 사용자를 기술적으로 보호하는 것을 정책의 목표로 삼는다. 다른 AI 에이전트의 활동을 감시하고 위험을 경고하며, 사용자의 최종 의사를 확인함으로써 피해를 예방하는 '수호자(Guardian)' 역할의 기술 개발과 보급이 필요하다. 이를테면 '옴부즈맨 에이전트' 핵심 기술이 개발되어야 하는데 이는 다른 AI 에이전트의 활동을 실시간으로 모니터링하여 사용자의 평소 패턴을 벗어나는 비정상적 행위를 탐지 및 경고하고 금융 거래, 개인 정보 전송 등 중요하며 되돌릴 수 없는 작업을 수행하기 전, 별도의 안전한 채널을 통해 사용자의 의사를 명확하게 재확인하는 개입 절차를 수행한다. 또 대화 내용, 요청 사항 등을 분석하여 피싱, 스캠, 사칭 등 사기 행위에서 나타나는 전형적인 패턴을 식별하고 사용자에게 고지하는 기능을 한다. 'AI 레드팀' 기술과 유사하게, 방어적 AI 기술 개발에 대한 R&D 투자가 필요하다.

'옴부즈맨 에이전트' 운영을 위한 기술 기반도 마련되어야 한다. 금융, 쇼핑 등 민감한 권한을 위임받는 AI 에이전트 사업자를 대상으로, 활동 기록(로그)을 표준화된 형식으로 생성하도록 의무화해야 한다. 유럽연합에는 고위험 AI 시스템에 대해 투명성과 감독 가능성을 요구하는 AI Act와 같은 AI 법이 있는데 이와 같이 외부 보안 에이전트와의 연동 채널 확보를 법제화할 필요가 있다.

AI가 제시하는 정보를 비판적으로 수용하고, 여러 관점의 정보를 비교하여 사실을 확인하는 능력을 함양하는 교육을 강화할 필요도 있다. 정서형 AI를 건강하게 사용하는 방법과 '옴부즈맨 에이전트'의 경고를 해석하고 대처하는 방법에 대한 교육 콘텐츠를 개발하고 보급해야 한다.

이런 정책을 시행하기 위해 2025~2026년에는 '옴부즈맨 에이전트'의 이상을 탐지하고 사용자 의도를 재확인 등 핵심 기술 R&D에 집중해야 한다. 또 2027~2028에는 실증 및 표준화 단계가 이뤄져야 하는데, 규제 샌드박스를 활용한 기술 실증, AI 에이전트 활동 기록 표준화를 추진하는 단계이다. 2029년 이후, 주요 플랫폼을 중심으로 기술 보급을 유도하고, 국제 표준화 협력을 추진하여 보급과 확산이 이뤄지도록 한다.

– 임화섭(한국과학기술연구원 단장)

AI 생성 이미지

5. 노동 변화 : 사라지는 직업, 재설계되는 노동

일하는 방식 변화와 생산성 향상

AI 에이전트는 많은 직종에서 일하는 방식 변화와 생산성 향상을 이끌 것으로 예견된다. AI 에이전트는 근로자의 능력 확장(augmentation) 도구로서 노동 생산성과 접근성을 크게 증진할 것이라 기대된다. AI 에이전트로 인해 근로자의 숙련성이 보다 빨리 노후화될 것이라 예상되므로 이에 대비한 지속적인 재교육과 업스킬링이 요구된다.

일부 직종에서 AI 에이전트가 노동력을 대체할 것이라는 예상도 할 수 있다. AI 에이전트로 인한 노동생산성의 비약적 상승은 노동 수요를 떨어트리고 임금 압박을 가져올 수도 있다. 골드만삭스는 AI가 정규직 일자리 3억 개를 대체할 수 있다고 보고하였고 한국개발연구원은 2030년에는 국내 일자리 열 개 가운데 아홉 개에서 AI가 70% 이상의 업무를 대체할 수 있다고 보고했다.

이제까지도 기술 혁신은 늘 노동시장의 변화를 이끌어 왔다. 하지만 AI 에이전트는 전문직을 포함한 광범위한 직업군에 노동 변화를 초래한다는 특성이 있다. 증기기관, 전기, 정보통신기술 등 과거의 기술 혁신은 주로 블루칼라 일자리와 높은 교육 수준을 요구하지 않는 직업군의 노동을 대체하는 경향을 보였으나, AI 에이전트는 높은 교육 수준과 전문성이 요구되는 직업군의 노동을 대체할 가능성이 크다는 차이가 있다.

조지타운 대학의 CSET(Cneter for Security and Emerging Technology)에 따르면 회계, 감사, 법무 등의 전문직이 AI 노출도가 높으면서 AI 보완성은 낮아 가장 취약한 직업군이다(〈그림 1〉 참조). 산업연구원은 AI가 전체 일자리의 13.1%를 대체할 것이며, 대체 일자리의 약 60%가 전문가 직종인 것으로 예측하는 자료를 내놓았다.

노동 변화 관련 기술의 긍정적 영향

AI 에이전트는 반복적이고 시간이 많이 소요되는 작업을 자동화하여 근로자들이 고차원적인 업무에 집중할 수 있도록 함으로써 업무 생산성을 높인다. 일례로 골드만삭스, 씨티그룹, 모건스탠리 등 글로벌 금융사들은 AI 에이전트 도구를 전사적으로 도입했으며, 모든 직원이 문서 요약, 콘텐츠 초안 작성, 데이터 분석 등 다양한 업무에서 AI 에이전트의 도움을 받도록 하여 업무 생산성을 높이고 있다.

AI 에이전트는 근로자의 능력을 단순히 대체하는 것이 아니라 이를 확장하는 방식으로 업무 생산성을 향상시킨다. AI 에이전트는 정보 처리 및 의사 결정 지원 등을 통해 근로자의 능력을 확장하여 업무 생산성을 높이는 것이다.

AI 에이전트의 능력 확장 기능은 다양한 인재가 노동시장에 진입할 수 있도록 장려한다. AI 에이전트를 통한 언어 번역, 시청각 보조는 장애인이나 외국인 근로자의 업무 참여를 어렵게 하던 장애물을 없애주기 때문이다.

AI 에이전트를 통한 업무 능력이 증대되면서 인구 감소 때문에 야기된 노동력 부족이 어느 정도 메꿔질 수 있다. AI 에이전트는 다수 근로자를 필요로 하던 업무를 1인이나 소수 근로자가 수행할 수 있도록 함으로써, 노동력 부족을 덜어주는 동시에 1인 기업 등의 확대를 촉진하게 된다.

AI 에이전트의 작업 자동화와 능력 확장 기능은 개인이나 소수 인원의

창업을 촉진한다. AI 에이전트에 능숙한 인력은 이를 이용하여 다양한 지식과 스킬이 요구되는 업무를 효율적으로 수행할 수 있어, AI 에이전트는 창업에 필요한 비용을 크게 감축하기도 한다. AI 에이전트는 특히 상대적으로 소규모이며 업무 조직화 정도가 낮고, 만성적인 인력 부족을 겪는 시민단체나 협동조합, 사회적기업 등의 영역에 큰 혜택을 제공할 것으로 보인다.

노동 변화 관련 기술의 부정적 영향

AI 에이전트가 업무 보조를 넘어 노동력을 대체하면서, 일부 직종에서는 대규모 구조적 실업이 발생할 가능성이 커진다. 특히 텔레마케터, 번역/통역사, 회계사, 감사, 법무 보조인, 프로그래머 등의 일자리가 빠른 속도로 감소할 것으로 보인다. AI 에이전트가 그 자리를 대체할 가능성이 큰 직종에 교육 수준과 전문성이 높은 직종이 다수 포함되어, 기존 일자리 정책의 변화가 요구된다.

최근 마이크로소프트가 발표한, AI로 인해 위험에 처한 직종에는 역사가, 번역가, 데이터 과학자, 수학자, 에디터 등 대학 이상의 고등교육을 요구하던 전문직이 다수 포함되어 있다. 이에 AI 에이전트에 의해 일자리가 위협받는 직종 종사자들의 저항으로 사회적 갈등이 유발될 가능성도 있다. 최근 '타다'를 둘러싼 택시업계의 저항이나 원격 진료에 대한 의료계의 저항은 AI 에이전트가 일부 직종 종사자의 저항을 불러올 수 있음을 시사한다.

생성형 AI 에이전트의 도입으로 거의 모든 직군에서 근로자가 기존에 지닌 기술의 노후화가 급격하게 진행될 것으로 예상된다. AI 에이전트의 도입으로 언어, 전산, 자료 수집 및 분석 등 근로자가 축적한 기존의 기술이 빠른 속도로 노후화되고 대체되고 있다. 따라서 AI 에이전트를 활용한 리스킬·업스킬 증진을 위한 재교육이 필요하고 대상 근로자 수가 크게 늘 것으로 보인다.

AI 에이전트를 잘 활용하는 인력과 그렇지 못한 인력 간에 생산성 격차가 커지며, 직업 안정성이나 임금 격차가 심화될 가능성이 커질 수 있다. AI 도구가 전문 지식을 갖춘 고숙련 근로자의 창의력을 향상시켜 이들의 업무 생산성을 크게 증진한 반면 저숙련 근로자의 업무 생산성 향상은 미미했다는 연구 결과는 AI 에이전트로 인한 생산성과 격차 심화를 시사하고 있다.

정책으로서 재교육과 평생 학습 강화 필요

근로자 재교육 강화와 일부 직종의 출구 전략 마련이 필요하다. 과거 기술 변화가 주로 블루칼라 일자리와 저숙련 일자리를 대체했다면, AI 에이전트는 전문직을 포함한 광범위한 직업군에 영향을 줄 것이 예상되어 근로자 재교육과 성인 평생학습에 대한 수요의 폭이 크게 확대되고 있다. 기술 변화로 인한 구조적 실업이 예상되는 일부 산업이나 직종의 근로자에게 재교육이나 직무 전환 등을 통해 새로운 업무와 직종에 적응할 수 있도록 유도하는 정책이 필요하다.

정책 대안으로, 우선 AI 에이전트 관련 직종 근로자에게 재교육과 업스킬링 기회를 제공하기 위해, 기업과 근로자가 선택할 수 있는 기존 교육 프로그램을 확대해야 한다. 현재 정부는 스마트직업훈련플랫폼(STEP) 사업을 통해 재직자, 구직자를 대상으로 AI 관련 과목을 포함한 온라인 직무 교육을 대부분 무료로 제공 중이지만 교육 시간이 과목당 열 시간 내외로 짧아 심화 교육을 제공하는 데는 한계가 있다.

단기적으로 현재의 근로자 재교육 프로그램을 개편하여, 다양한 AI 에이전트를 능숙하게 다룰 수 있을 정도로 심화된 근로자 재교육을 제공하도록 재정을 지원해야 한다. 중기적으로 AI 에이전트가 전문직에 큰 영향을 주기 때문에 전국의 대학(원)이 더 다양하고 전문성 높은 AI 직무 재교육을 제공할 수 있도록 (가칭) 'AI 직무 재교육 사업'을 지원해야 한다. 일례로 싱가포르의 SkillsFuture 프로그램을 들 수 있는데 이는 국공립 대학(Autonomous Universities), Polytechnics, 기술교육기관(Institute of Technical Education) 등에서 기술 재교육을 받을 때 드는 비용을 최대 90%까지 지원하며, 이들 다양한 기관을 통해 광범위한 AI 관련 교육 프로그램을 제공하고 있다. 그리고 정부는 프로그램을 인증하는 역할을 한다.

근로자 재교육이나 업무 전환은 개인 근로자뿐 아니라 개별 기업에도 부담이 크므로, 지속적인 고용 유지 차원에서 기업을 지원할 필요가 있다. 근로자 재교육과 직무 전환이 적극적으로 이루어지기 위해서는 이들의 재교육과 직무 전환으로 인한 기업의 부담을 경감시킬 수 있도록, 중소사업체 위주로 세제 혜택 등을 통한 지원을 제공하는 것도 고려할 만하다. 싱가포르의 SkillsFuture 프로그램은 개인뿐 아니라 중소 사업체에도 근로자 교육 프로그램 이수나 직무 전환 때문에 발생하는 비용을 지원하고 있다.

일부 직종에서 AI 에이전트로 인한 구조적 실업이나 산업 대체가 불가피하므로, 이에 대한 선제적 정책 개입이 필요하다. AI 에이전트에 의해 일자리를 위협받는 직종에 근무하던 근로자의 저항으로 사회적 갈등이 커지기 전에, AI 에이전트 기술의 사회 영향 평가를 통해 일자리가 위협받을 가능성이 큰 산업이나 직종에 어떤 방식의 출구 전략을 제공할지를 선제적으로 논의하고 대비해야 한다.

AI 에이전트의 기능을 활용한 창업을 지원하여 새로운 일자리를 창출하고 경제·사회 혁신 활성화를 꾀해야 한다. AI 에이전트의 작업 자동화와 능력 확장 기능을 활용한 1인 기업 창업을 지원할 필요가 있다. 중소벤처기업부의 '초격차 스타트업 1000+ 프로젝트' 같은 기존 기술 기반 스타트업 지원 프로그램은 AI 에이전트를 이용한 알고리즘 기업의 경우 1인 기업이 될 수 있다는 점을 고려하여 지원해야 한다.

– 박희제(경희대학교 교수)

6. 노동 대전환
: AI 오케스트레이션 교육·인증으로 '1인 창업' 활성화

생성형·에이전틱(Agentic) AI는 단일 과업을 보조하는 차원을 넘어 복합 의사 결정·실행 자동화의 다중 에이전트 협업 단계로 진입하고 있다. 이에 따라 사용자는 단순히 명령을 내리는 역할에서 벗어나, 여러 AI 에이전트의 작업을 지휘하고 조율하는 '오케스트레이터'로 전환되고 있다. 이러한 변화는 '알고리드믹 앙트레프레너십(Algorithmic Entrepreneurship)'이라는 새로운 경제 패러다임을 부상시키고 있다. 이는 개인이 AI 에이전트를 활용해 비즈니스 아이디어 구상부터 조직 설계, 제품 판매에 이르는 전 과정을 자동화하며 새로운 사업 기회를 창출하는 것을 의미한다. 마이크로소프트 등 주요 기업들은 시퀀셜(Sequential), 콘커런트(Concurrent), 그룹챗(Group Chat) 등 오케스트레이션 패턴을 아키텍처 가이드로 정립하며 현장 적용을 확산시키고 있다. 이는 곧 개인의 아이디어가 최소한의 자원으로 신속하게 사업화될 수 있는 기술적 기반이 마련되었음을 시사한다.

노동 전환 관련 기술의 긍정적 영향

조사, 문서 작성, 코딩, 고객 응대 등 반복적이거나 분석적인 과업이 자동화되면서 성과와 속도는 높아지고 오류는 감소하고 있다. 이를 체계적으로 지원하기 위해 표준 커리큘럼(Lv1~3)을 개발하여 기술 내재화를 돕고, 국가 실습 플랫폼에서 ROI(투자수익률)/KPI(핵심성과지표) 대시보드를 제공해 성과를 시각적으로 관리하도록 지원이 필요하다.

다중 에이전트 오케스트레이션은 소수 인력, 심지어 1인만으로도 과거 대규모 조직이 수행하던 복잡한 업무를 처리할 수 있게 한다. 오픈AI의 CEO 샘 알트만은 이러한 변화를 근거로 AI를 통해 '1인 유니콘 기업'의 등장이 가능할 것이라고 예측한 바 있다. 이는 한 명의 창업가가 다수의 AI 에이전트를 관리하며 수천억 원의 가치를 지닌 기업을 만들 수 있다는 의미이다.

이러한 생태계에서는 특정 목적을 위해 실시간으로 생성되었다가 사라지는 '1회용 앱(Disposable Apps)'과 같은 새로운 형태의 비즈니스 모델이 활성화될 수 있다. 또한 영국의 법률 서비스 분야에서 AI가 계약서 검토, 법률 리서치 등을 자동화하여 서비스 효율을 극대화하는 것처럼, 고도의 전문성이 필요했던 영역에서도 알고리드믹 앙트레프레너십이 확산될 것이다. 이러한 흐름을 국가적 기회로 활용하기 위해, 교육 → 창업경진대회 → 마이크로 투자·바우처 → 클라우드 크레딧 → 공공 시범 구매와 같이 교육부터 사업화까지 전 단계를 지원하는 원스톱 '1인 창업 패스'의 구축이 필요하다.

초기 연구에 따르면 AI 기술을 도입할 때 비숙련자의 성과 향상 폭이 숙련자보다 더 크게 나타나고 있다. 이에 취약 계층을 대상으로 우선 바우처와 무상 교육 과정을 제공하여 기술 접근성의 진입 장벽을 낮출 필요가 있다.

노동 전환 관련 기술의 부정적 영향

고도의 오케스트레이션 역량을 지닌 소수와 AI에 의해 대체되거나 단순 보조 역할을 맡는 다수 간의 격차가 심화될 수 있다. 이는 마이크로 크

레덴셜(Lv1~3)과 전환 바우처 제도를 통해 지속적인 상향 이동 경로를 제공함으로써 해결할 수 있다.

기술 역전(Skills Inversion)도 우려되는데 고숙련 직무의 핵심 역량인 분석과 판단까지 자동화되면서, 오히려 여러 기술을 조율하고 관리하는 메타 스킬의 가치가 급상승하는 기대를 할 수 있다. 이는 교육 과정 내에서 결과물을 검증하고(Maker-Checker), AI 거버넌스 및 ROI를 관리하는 모듈을 강화하는 방법으로 해결점을 찾아갈 수 있다.

프로젝트 기반의 프리랜서 고용 형태가 확산되면서 사회보험의 사각지대에 놓이는 노동자가 증가할 수도 있다. 이 문제 해결을 위해서는 프로젝트 소득에 기반한 개인계정형 사회보장 제도인 '포터블 베네핏(Portable Benefits)'의 법제화를 연구하고 시범 사업을 추진해야 한다.

정책 수요와 대안

'오케스트레이션 교육·인증 – 창업·일자리' 선순환 구축으로 알고리드믹 앙트레프레너십 활성화할 필요가 있다. 초·중·고·대학·재직자를 대상으로 업무 분해, 프롬프트 엔지니어링, 도구 연동, 패턴 실습, 평가, 윤리를 포함하는 표준 교육과정 개발하고 국가 공용 실습 플랫폼(SaaS): 다중 에이전트 설계, 비용·안전성 평가, 케이스 스터디 및 API 라이브러리를 제공하는 클라우드 기반 실습 환경을 구축해야 한다. 역량 수준에 따라 Lv1(단순 도구 사용)부터 Lv3(거버넌스&ROI 설계)까지 구분하고, 프로젝트 기반으로 평가하여 실무 역량 인증이나 국가 자격(마이크로 크레덴셜)을 부여할 수도 있다.

교육 이수자를 대상으로 창업 아이디어 경진대회를 개최하여 우수 아이디어를 발굴하고, 마이크로 투자, 정액 바우처, 클라우드 크레딧, 공공 시범 구매 기회를 패키지로 제공하는 것도 고려할 만하다. 공공데이터 활용이나 가벼운 서비스 테스트에 대해 간소화된 절차를 적용하여 아이디어를 시장에서 빠르게 검증하도록 지원할 수 있다. 프로젝트 소득 기반의 개인계정형 사회보장 제도를 연구하고 단계적으로 도입할 필요가 있다.

– 임화섭(한국과학기술연구원 단장)

AI 생성 이미지

7. AI에이전트의 사회적 수용성 증대
: AI를 받아들이는 사회, 배제되는 사람들

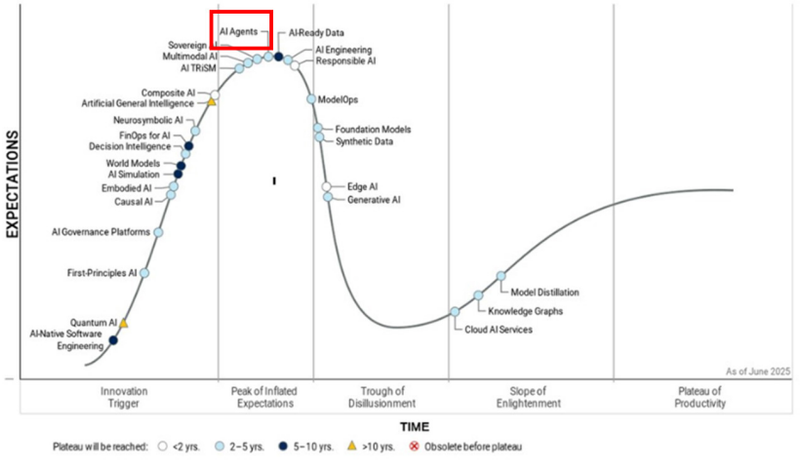
AI 에이전트의 사회적 수용성 증대 필요성

AI 에이전트는 지식 전달, 정보 탐색, 의사 결정 지원, 교육 혁신의 주체로 빠르게 자리 잡고 있으며, 이러한 변화는 사회 전반의 수용성 문제와 직결되고 있다. 고성능 AI 에이전트가 확산되면 사회 구성원 간 정보 접근성과 기술 이해도에 따라 혜택의 불균형을 심화시킬 수 있다. 이에 따라 AI 에이전트 교육 방식이 근본적으로 변화되어야 하며, 이는 미래 인력 양성과 핵심 역량 개발을 위한 새로운 패러다임으로 이어질 것이다.

AI 에이전트 성능이나 접근성 차이로 인해 사회경제적 불평등이 심화될 가능성이 있으며, 고성능 AI를 사용할 수 있는 계층과 그렇지 못한 계층 간 정보와 교육 격차가 커질 수도 있다. 특히 노년층, 장애인, 디지털 취약 계층이 배제될 위험이 크므로 이를 완화하기 위한 포용적 정책이 필요하다.

사회 전반에 AI 에이전트를 맹목적으로 신뢰하는 경향이 커질 경우, 의료, 법률, 소비 등 다양한 영역에서 부작용을 초래할 가능성이 생긴다. AI 에이전트에 대한 정확한 이해가 부족한 상황에서 AI 상담에 과도하게 의존하여 병원 진료를 기피하거나, AI 추천에 따른 과소비가 발생할 수도 있다. 실제로 2023년 미국 뉴욕에서는 변호사가 챗GPT를 활용해 작성한 소송 문건에 실제 존재하지 않는 가짜 판례가 포함되어 법정에서 문제가 된 사례가 있다.

〈Gartner Hype Cycle Identifies Top AI Innovations in 2025〉



성별·연령·신체 조건 등 개인별 특성에 따라 AI 에이전트 수용 경험이 달라지며, 이는 수혜 불균등으로 이어질 수 있다. AI 에이전트와 함께 성장한 세대는 의존성이 높고 멀티태스킹 능력이 향상하는 등의 특성을 보일 것으로 예상되며, 세대 간 인식과 활용의 차이가 사회적 갈등으로 이어질 우려도 있다.

기존 지식 전달 중심의 교육 체계는 AI 보편화 시대에 한계를 보이게 된다. AI 에이전트에 의한 노동 방식 변혁과 지식 구조의 전환에 대응하기 위해, 비판적 사고, 창의적 문제 해결, AI 리터러시 등 핵심 역량(core skill) 중심 교육으로의 전환이 필요하다.

사회적 수용성을 고려한 AI 에이전트 기술의 현황 및 전망

대규모 언어모델(LLM)로 대표되는 인공지능 기술 혁신으로 AI 에이전트가 사회 내에서 일상적 생활 방식을 변화시키는 주체로 떠오르고 있다. 이에 포브스는 AI 에이전트가 단순한 도구를 넘어 자율적 대리 행동 수행자로서 복잡한 일상 업무를 처리할 것으로 전망한 바 있다. 마이크로소프트는 AI 에이전트가 업무 생산성을 높이는 핵심 동력이 될 것으로 내다보며, 일상에서 개인의 생활 방식 전반에도 영향을 미칠 것이라고 분석했다.

현재 AI 기술의 활용이 일부 계층에 국한될 우려가 있으며, 이를 완화하기 위해 포용적 AI 기술을 개발하고 서비스 접근성을 개선하는 방안이 제안되었다. 마이크로소프트는 AI 사용자 인터페이스를 설계할 때 단순 시각적 표현을 넘어 인지적 다양성, 학습 스타일, 사회적 편견을 고려한 포용적 기술을 적용했다. 소외 계층이나 소수 커뮤니티의 의견을 반영할 수 있는 참여 기반 의사 결정 플랫폼인 '인클루시브.AI'와 같은 포용형 AI 서비스 기술이 개발되고 있다. AI 에이전트가 장애인 같은 특정 대상에게 맞춤형으로 제공될 수 있도록 사용자 편의성과 접근성을 고려한 UI/UX 디자인뿐만 아니라 다양한 대리 서비스를 제공하는 기술 개발과 정책적 지원 방안을 마련하는 것이 절실하다.

인공지능 거버넌스의 일환으로 사회적 수용성을 고려한 규제와 표준화가 국제적으로 추진되고 있다. 2024년 유럽평의회는 법적 구속력을 갖는 최초의 국제 AI 규범인 「Framework Convention on Artificial Intelligence(AI 기본조약)」를 채택했으며, ISO/IEC 42001은 접근성과 포용성을 포함한 AI 경영시스템 기준을 제시하고 있다.

한편, 각국은 디지털 리터러시 교육을 중심으로 관련 공공 정책을 시행하고 있으나, AI 에이전트의 자율적 대리 행위와 사회적 영향에 특화된 정책은 아직 충분하지 않은 상황이다. 유네스코의 생성형 AI 가이드와 국내의 AI 디지털교과서·취약 계층 대상 교육 정책은 이러한 공백을 보완하기

위한 초기 단계의 대응으로 평가된다.

향후 다양한 세대와 계층을 대신해 행동하는 AI 에이전트의 활용이 확산됨에 따라, 사회적 수용성을 제고하기 위한 맞춤형 에이전트 기술 개발과 이를 뒷받침할 제도·정책의 중요성이 더욱 커질 것으로 예상된다. 이와 함께 EU 인공지능법(Artificial Intelligence Act)과 같이 신뢰성과 사회적 수용성을 반영한 AI 인증제가 강화되어, 공공 및 민간 영역에서 도입 시 필수 기준으로 자리 잡을 가능성이 크다.

AI 에이전트의 사회적 수용성 관련 기술의 긍정적 영향

AI 에이전트는 법률·의학·금융 분야에서 공공용 AI 에이전트를 통한 상담과 정보를 제공하는 등 비전문가도 전문 지식에 접근할 수 있도록 하여 지식 민주화를 촉진하고 있다. 공공 플랫폼을 통한 동등한 정보 접근 기회 확대는 사회적 평등에도 기여하고 있다.

AI 에이전트를 활용한 맞춤형 학습, 토론과 탐구를 기반으로 한 교육의 확산을 통해 기존 교육 한계를 극복할 수 있다. 비판적 사고, 협업 능력, AI 리터러시 등 미래 사회 핵심 역량을 체계적으로 강화할 수도 있다.

다양한 연령, 성별, 장애 특성을 고려한 맞춤형 AI가 개발되면 포용적 사회 구현에 기여할 것이다. 디지털 소외 계층 대상으로 한 AI 활용 교육은 사회 통합을 촉진하고 정보 불평등을 완화할 것이라 기대된다.

AI 에이전트의 사회적 수용성 관련 기술의 부정적 영향

고성능 AI 에이전트에 대한 접근권이 제한될 경우, 사회경제적 불평등이 심화될 수 있다. 충분한 사회적 지원이 부족할 경우, 경제력이 낮은 집단이 최신 AI 기술 혜택에서 배제될 가능성이 지속적으로 제기되어 왔다.

기술 이해도와 활용 능력 차이로 인한 세대와 계층 간 격차가 심화해지는 것도 우려된다.

AI의 정보 오류와 편향으로 인해 진료 기피, 소비 패턴 왜곡, 정치 사회적 편향 심화 등 잘못된 의사 결정이 발생할 수도 있다. AI를 전문가로 맹목적으로 신뢰할 경우, 비판적 사고가 약화되고 사회적 취약성이 증대될 가능성도 있다.

AI 에이전트 활용 경험이 축적된 세대와 그렇지 않은 세대 간 인식과 활용 차이로 갈등이 심화될 수도 있다. 국내에서는 디지털 교과서 도입에 대한 법적 혼선과 학교 현장의 운영 혼란 사례가 보고된 적이 있다. AI 의존 세대는 높은 멀티태스킹 능력을 보이는 동시에 집중력이 낮아지고 AI에 대한 의존성이 높아질 수 있어 전통적 교육 성과가 떨어질 가능성이 있다.

정책 수요와 대안

AI 에이전트를 활용해 비전문가의 지식 격차를 해소하고 사회 전반의 정보 접근성을 제고할 필요가 있다. 이를 위해 단기적으로 의학·법률·금융 등 전문 영역에서 신뢰 기반의 지식을 제공하는 AI 에이전트를 개발하는 사업을 추진해야 한다. 예를 들어 '신뢰 가능한 지식 제공 AI 에이전트 기술 개발 사업'과 같은 연구 과제를 통해, 전문 지식을 비전문가가 이해할 수 있도록 상식 수준의 설명과 부가 정보를 제공하는 기술을 개발할 수 있으며, 이는 기존 데이터와 기술을 활용해 단기 정책 과제로 추진할 수 있다.

아울러 공공 AI 지식 에이전트 플랫폼을 구축해 국민의 정보 접근성을 높일 필요가 있다. 정부 기관, 공공 연구기관, 학회 등의 신뢰 가능한 콘텐

츠를 연계한 AI 에이전트 기반 국민 지식 도우미를 모바일·웹·공공 키오스크 등 다양한 채널로 제공함으로써, 공공 데이터와 공공 LLM을 활용한 단기 정책 추진이 가능하다.

한편 AI 에이전트 접근성과 기술 이해도 격차로 인한 사회경제적 불평등 확산을 방지하기 위해, 중기적으로는 성별·연령·장애 등 다양성을 반영한 맞춤형 AI 에이전트 개발 사업을 추진할 필요가 있다. '사용자 다양성 기반 인클루시브 AI 에이전트 기술 개발'과 같은 사용자 편의성과 접근성을 고려한 UI/UX를 제공하고, 개인의 지식 수준과 생활 환경, 신체적 특성에 맞춰 대리 역할을 수행하는 AI 에이전트 개발이 요구된다. 이와 함께 AI 사용자 경험 설계 고도화 추세를 고려해 관련 기술 분야에 대한 투자를 중기적으로 단계적으로 확대하는 정책을 추진하는 것이 바람직하다.

또한 단기적으로 디지털 소외 계층을 대상으로 'AI 활용 시민 리터러시 교육 바우처'를 도입해 AI 에이전트 활용 교육 기회를 제공할 필요가 있다. 과기정통부와 고용노동부가 협업해 지자체 연계 평생교육센터를 확대 운영하는 방안도 고려할 수 있다.

중기적으로는 AI 접근 취약 계층을 위한 'AI 에이전트 공공 접근 보장법' 제정을 통해 AI 접근에 대한 차별 해소를 추진하고, 향후 인공지능 기본법 등 관련 법제에 단계적으로 확대 적용하는 것이 바람직하다.

– 강동오(한국전자통신연구원 책임연구원)

8. 사회적 수용성 : AI를 신뢰할 수 있는 사회

AI 에이전트의 사회적 수용성 증대 필요성

AI가 우리 일상생활 깊숙이 스며듦에 따라 AI 에이전트의 사회적 수용성과 공평한 접근성 확보가 그 어느 때보다 중요해졌다. AI는 사회 격차를 확대하거나 좁힐 수 있는 양날의 도구로서, 포용적 교육·정책·투명성 등 종합적 노력이 뒷받침되지 않으면 또 다른 불평등을 불러올 우려가 있다.

사회 구성원의 AI 시스템 신뢰도는 지역별·계층별로 편차가 크며, 선진국에서는 AI 신뢰도가 39%에 불과한 반면 신흥국에서는 57%에 달한다. 이는 AI 수용도에도 영향을 미쳐 신흥국은 84%, 선진국은 65%의 AI 수용도를 보이고 있다. AI에 대한 신뢰는 기술 도입의 핵심 요소로, 신뢰 부족은 AI 채택을 막고 사회 전반의 AI 활용을 제한할 수 있다.

AI는 많은 이에게 효익을 주지만 개인의 능력·접근성·학습 기회·경제력에 따라 혜택의 격차가 커져 오히려 기존 사회·경제적 격차를 심화시킬 수 있다. 도시 고학력층 위주로 AI 도구와 데이터 활용이 확산되고 농촌 등의 취약 계층은 활용이 저조하여 학습 기회와 정보 접근 격차가 오히려 벌어지고 있다.

교육 분야에서 AI 도입은 유리한 환경의 학교에서 앞서 나타나고 있다. 미국의 한 조사에서 교외의 부유한 학군이 도시·농촌의 저소득 학군보다 AI 활용에서 앞서 있다는 초기 징후가 포착되었다. 교사의 AI 활용 역량과 지원 수준 차이로 학교별로 도입 편차가 발생하고 있으며 일부 교사들은 AI에 대한 두려움으로 AI 활용을 기피하는 실정이다. 그런데 교사의 AI

활용 기피는 교실 내 기술 수용 격차를 더욱 심화시키는 결과를 낳는다.

사회적 수용성 관련 기술의 긍정적 영향

AI 에이전트를 책임 있게 활용할 경우 약해진 사회 신뢰를 강화할 수 있는데, 사회적 신뢰 향상은 경제적 성장에도 기여하게 된다. 한 국가 안에서 신뢰하는 사람의 비율이 10%포인트 증가하면 연간 1인당 실질 GDP 성장률이 약 0.5%포인트 상승한다는 보고도 있다.

AI가 왜 그런 결정을 내렸는지 어떻게 동작하는지 설명이 가능하고, 같은 상황에서 일관되게 행동하는, 이른바 투명한 AI 시스템은 사용자에게 신뢰감을 주어 수용성을 높이고, 장기적으로는 사람과 AI 에이전트 사이에 신뢰 관계를 구축하게 된다.

AI 기술은 시각장애인을 위한 이미지 설명, 청각장애인을 위한 자동 자막 등 유용한 기능으로 장애인의 정보 접근성을 높여주어, 기술 활용에 따라 일부 디지털 격차 해소를 기대하게 한다. AI는 사용자 수준에 맞춘 교육 도구를 제공하여 기술 소외 지역의 정보 접근을 개선하고, 디지털 전환의 혜택을 사회 전반에 고르게 확산시킬 수 있다.

AI를 활용한 맞춤형 학습은 학생 개개인의 수준과 학습 스타일에 맞게 콘텐츠를 제공하여 이해도를 높이고 학습 격차를 완화하는 데 도움이 된다. AI 기반 교육 플랫폼은 지리적·사회경제적 한계를 넘어 전 세계 학생들에게 양질의 교육 자료를 제공함으로써 교육 접근성 불균형을 해소하도록 해준다.

사회적 수용성 관련 기술의 부정적 영향

AI 에이전트의 결정 과정이 불투명하거나 '의도'를 알 수 없으면 이용자들이 시스템을 불신하게 되어 수용도가 낮아진다. 딥페이크 영상과 자동화된 금융 사기 등 AI 악용이 증가하면 사회적 신뢰 기반이 흔들릴 수 있으며 사회 전반의 불신이 고조될 우려도 있다.

AI 활용 능력의 차이에 따라 개인 간 성과 격차가 기하급수적으로 커질 수 있으며, 기술이 기존 격차를 확대하고 고착화시킬 위험도 있다. AI 기반 교육과 정보 활용은 도시 엘리트 계층에 집중되고 농촌·저소득층은 소외되는 경향이 있다. 따라서 AI 도입이 학습 기회와 정보 접근 격차를 더욱 심화시키고 있다는 우려도 제기되고 있다.

부유한 학군의 학교들은 AI 도입 및 활용에 적극적인 반면 재정이 열악한 학교들은 준비 부족으로 출발선부터 격차가 벌어지는 양상을 보인다. 일부 학교는 AI 사용을 금지하거나 AI 활용에 대한 교사의 저항이 있는데 이는 학생의 학습 환경 격차와 교육 성과 불평등을 심화시킬 위험으로 이어질 수도 있다.

정책 수요와 대안

AI 에이전트에 대한 국민 이해도를 높이고 AI 에이전트를 확산시킬 기반을 마련하며 책임 소재 명확화를 위한 제도를 마련해야 한다. 전 국민 대상 AI 리터러시 교육을 확대하는 동시에 다양화하고 고령층, 장애인, 다문화 가정 등 취약 계층 대상 디지털 역량 강화 교육인 '디지털배움터' 사업도 전 국민 대상 AI 교육으로 확대할 필요가 있다.

AI 바우처를 발급하여 국민 모두가 사각지대 없이 AI 서비스를 활용할 수 있도록 지원하는 한편 AI 바우처를 통한 AI 디지털 리터러시 교육을 연계할 필요도 있다. AI 바우처를 통해 AI 유료서비스를 이용하려면 필수적으로 AI 리터러시 교육을 이수하도록 하고, 교육 이수 시 디지털 지갑

에 보관 가능한 디지털 배지를 발급하는 방법도 있다.

공공 웹사이트나 서비스에 적용할 AI 접근성 표준 개발 및 공공 서비스 AI 에이전트 적용을 확대해야 한다. 공공 웹사이트나 서비스에 적용할 AI 접근성 표준을 개발하고, 취약 계층 친화적 AI 인터페이스 개발도 고려할 만하다. 공공 AI 에이전트를 신속히 도입하기 위한 공공 MCP 서버 허브, 공공 MCP 마켓 플레이스 등을 구축하여 AI 에이전트 도입 기반을 마련할 수 있다. 이는 과기정통부에서 2025년 추진 중인 DPG 허브 2차 사업 범위에 MCP 서버 허브 및 공공 MCP 마켓 플렛이스를 구축하는 것을 포함한다.

AI기본법 시행령에 AI 에이전트의 법적 지위와 책임 소재를 명확히 규정하는 내용을 포함하여 불확실성을 해소해야 한다. AI 에이전트가 잘못된 의사 결정을 했을 때 사용자, 개발자, 플랫폼 중 명확한 책임 주체 및 배상 체계 등을 법률로써 명시해야 한다. 위험성이 높은 분야의 AI 에이전트에 대해서 EU 사례를 참고하여 강화된 안전성 기준과 감독 체계를 마련하여 적용할 필요가 있다.

과기정통부는 2022년부터 AI 서비스 확산에 따른 사회적 영향을 살펴보고 사전 대응하기 위해 'AI 서비스 사회적 영향 평가'를 조사·분석하고 있는데 이렇게 정부, 기업, 전문가, 시민단체 등이 모여 AI 에이전트 도입에 따른 사회적 영향을 모니터링하고 갈등을 조정해야 한다. AI 에이전트 도입에 따라 발생하는 긍정적/부정적 영향을 조사하고 분석하며 AI 에이전트 도입 과정에서 발생하는 이해 관계 충돌을 선제적으로 해소해야 한다.

AI 에이전트의 신뢰성을 확보하고 AI 에이전트의 잘못된 의사 결정을 맹신하지 않도록 관련 기술을 개발하고 제도화를 추진해야 한다. AI 에이전트의 성능이나 안전성 인증 제도를 도입하고 AI 표시제를 의무화하고 AI 에이전트의 정보 신뢰도와 윤리성을 평가하는 기준을 마련하여 일정 수준을 충족한 서비스일 경우 '신뢰 가능 AI' 인증 마크를 부여하는 것도 한 방법이다. AI가 생성한 정보에 대해서는 AI 생성 정보 출처 표시를 의무화하는 것도 고려할 만하다.

AI 에이전트가 생성하는 정보의 사실 여부를 검증하고, 허위·조작 정보 탐지와 차단 기능을 갖춘 기술을 개발해야 한다. AI 에이전트가 생성한 콘텐츠의 신뢰도를 자동 분석하여 공신력 있는 데이터베이스와 대조하는 사실 확인 알고리즘 개발이 필요하며 AI를 악용해 생성한 딥페이크 영상, 가짜 뉴스, 조작 이미지 등을 식별하고 이를 차단하는 AI 오남용 탐지 시스템도 개발되어야 한다. 이의 일환으로 과기정통부는 2025년 '디지털 기반 사회 현안 해결 프로젝트' 사업을 통해 'AI 기반 아동·청소년 온라인 성 착취 선제적 대응 시스템 도입'을 추진 중이다.

AI 에이전트가 보편화함에 따라 AI 맞춤형 교육 생태계를 조성하고 AI 일상화 시대를 대비한 미래 인재 양성 체계가 구축되어야 한다. AI 기술을 활용한 교육 환경 구축과 교육 효과 극대화를 꾀할 수 있다. 교육기본법 개정을 통해 AI 활용 교육의 법적 근거를 마련하고 AI 디지털 교과서 확산과 교사 역량을 강화하는 프로그램을 체계화할 필요가 있다. AI 교육 플랫폼을 고도화하고 학습 효과를 분석하는 기술 개발도 요청된다.

AI 일상화 시대에 필요한 융합형 인재 양성을 위해 교육 과정을 혁신하고 인력 양성 체계를 구축해야 한다. 기존 법뿐만 아니라 교육기본법에 인공지능 교육과 윤리 조항을 신설하여 AI 윤리 교육과 리터러시 증진을 위한 관련 교육을 의무화할 필요도 있다. AI·융합 전공을 확대하고 산학 협력 기반 실무형 교육 과정를 도입하며 AI 교육 콘텐츠 개발과 교육 방법론 연구를 지원할 필요가 있다.

– 김태원 (한국지능정보사회진흥원 수석)

교육기본법 제23조(교육의 정보화) ① 국가와 지방자치단체는 정보화교육 및 정보통신매체를 이용한 교육을 지원하고 교육정보산업을 육성하는 등 교육의 정보화에 필요한 시책을 수립 · 실시하여야 한다.

② 제1항에 따른 정보화교육에는 정보통신매체를 이용하는 데 필요한 타인의 명예 · 생명 · 신체 및 재산상의 위해를 방지하기 위한 법적 · 윤리적 기준에 관한 교육이 포함되어야 한다.

제23조의4(인공지능 교육 및 윤리)(신설)

① 국가와 지방자치단체는 모든 국민이 인공지능 기술을 올바르게 이해하고 활용할 수 있도록 인공지능 리터러시 교육에 필요한 시책을 수립 · 실시하여야 한다.

② 국가와 지방자치단체는 인공지능 기술의 개발과 활용 과정에서 발생할 수 있는 윤리적 문제에 대응하기 위하여 인공지능 윤리 교육을 의무화하여야 한다.

③ 제1항과 제2항에 따른 교육에는 인공지능의 작동 원리, 사회적 영향, 편향성과 차별 방지, 개인정보 보호, 책임 있는 사용법 등이 포함되어야 한다.

AI 윤리 교육 및 AI 리터러시 증진을 위해 관련 교과목을 편성

9. 사회적 수용성과 미래 교육
: 신뢰받는 AI는 어떻게 만들어지는가

AI 에이전트는 단순한 질의 응답 시스템을 넘어 자율적 판단이나 행동이 가능한 지능형 시스템으로, 사회 전반의 업무 방식과 학습 환경에 직접 영향을 끼칠 수 있다. AI 에이전트 기술은 급속히 확산되고 있다. 미국, EU, 일본 등은 산업·공공·교육 현장에서 AI 에이전트의 활용이 빠르게 늘어나고 있으며, 2030년까지 전체 기업의 86%가 AI 자동화를 통해 비즈니스 운영이 확장될 전망이다. 이에 따라 사회 전반의 수용성을 높이기 위해 기술 융합 전략과 정책 지원이 뒷받침되어야 할 것이다. 사회적 수용성 확보는 기술 확산 속도와 방향을 좌우하며, 교육 분야는 장기적으로 AI 활용 인력의 질을 결정하는 핵심 영역이 될 것이다.

AI 에이전트는 정책·산업·교육 등 공공성이 높은 영역에 깊이 관여하고 신뢰성·투명성·책임성 확보가 사회적 수용성의 전제가 되어야 한다. 사회적 수용성이 높을수록 공공 부문과 민간 영역에서 AI 활용 범위와 속도가 빨라질 것으로 보인다. 그러나, 사회적 수용성 측면에서 프라이버시 침해, 일자리 대체 불안, AI 결정 불투명성, 세대별 격차 등이 해결 과제로 여전히 남아 있다. AI 에이전트에 대한 설명 가능성을 통한 기술 신뢰를 확보하고 맞춤형 상호 작용, 멀티모달이나 다언어 지원 등과 같이 접근성과 포용성을 높이는 기술 개발도 이뤄져야 한다.

AI 에이전트를 활용한 교육은 개별 맞춤·실시간 피드백·평생학습 지원을 통해 기존 교육 시스템의 한계를 보완할 수 있다. AI 에이전트 기술을

기반으로 개인화 학습을 지원하고, 지속적인 실시간 피드백을 통한 학습 성취도를 높이는 등 교육 품질 고도화를 꾀할 수 있다. 기술 발전 속도가 빠른 만큼, 교육 시스템이 AI 에이전트를 포함한 학습 생태계로 진화하지 않으면 국가 경쟁력 격차가 크게 벌어지는 등 우려할 만한 일이 발생할 수 있다.

사회적 수용성과 미래 교육 관련 기술의 긍정적 영향

AI 에이전트를 통한 민원 처리, 정책 분석, 시민 질의 응답이 즉시 가능해져 정부-시민 간 소통 속도와 품질이 향상될 수 있다. AI가 복잡한 정책 정보를 요약 또는 시각화해 일반 시민이 정책 논의에 쉽게 참여할 수 있게 하여 참여 민주주의 확대를 기대할 수 있다.

AI 에이전트는 언어, 문해력, 장애 여부와 관계없이 음성, 텍스트, 시각 자료 등 맞춤형 인터페이스로 정보를 제공한다. 덕분에 더 다양한 형태로 공공 서비스를 효율적으로 제공하며, 맞춤형 행정 서비스, 시민참여형 의사 결정 등도 가능해질 것으로 보인다.

고령층, 장애인, 농어촌 주민 등 상대적 디지털 소외 계층도 AI 에이전트를 통해 다양한 지능형 서비스를 제공받을 수 있고 이들을 대상으로 AI 에이전트가 맞춤형으로 금융, 교육, 의료 서비스를 받을 수 있도록 쉬운 접근법과 도움을 제공할 수도 있다.

개인별로 평생 맞춤형 학습을 제공하며, 교육의 품질 향상이나 교육받는 사람의 성취도를 높일 수 있다. AI 에이전트가 성취도, 흥미도, 학습 속도 등 학습자 데이터를 기반으로 맞춤형 학습 경로를 설계하며, 재직자·구직자를 대상으로 직무 맞춤형 AI 학습 코칭을 할 수도 있다. 채점이나 자료 정리 등 반복적인 행정 업무를 AI 에이전트가 자동으로 처리하고 교사는 창의적인 심화 교육에 집중하여 교육의 질을 향상시킬 수 있다.

사회적 수용성과 미래 교육 관련 기술의 부정적 영향

인터넷 등 네트워크, 디바이스와 같은 AI를 활용할 수 있는 환경이 취약한 계층은 오히려 불리해져 사회적 불평등이 심화될 우려가 있다. 불완전하거나 편향된 AI 에이전트의 응답은 이러한 취약 계층이 잘못된 결정을 내리게 유도할 가능성도 지니고 있다. AI 에이전트 기반 학습 도구에 접근하기 쉬운 교육기관과 그렇지 않은 기관 간의 격차가 확대될 수도 있다.

AI 에이전트가 내린 판단 근거가 불분명할 경우, 정부 행정이나 정책에 대한 신뢰도가 떨어지기도 한다. 또한 AI 에이전트가 의사 결정 과정에 깊이 관여한 후 오류가 발생했을 때 법적·윤리적 책임 주체가 불명확하여 신뢰도를 더 빨리 떨어트릴 수 있다.

국민 일상과 교육 현장에서 AI 에이전트에 답을 의존하면 비판적 사고나 문제 해결 능력이 저하될 우려가 있다. 문제를 스스로 분석하고 해결 전략을 설계하는 능력, 정보 탐색 능력과 논리적 추론 과정, 창의적 발상 등이 퇴보하여 사고가 획일화할 위험도 있다.

정책 수요와 대안

AI 동작에 대한 불확실성, 오작동 우려, 잘못된 응답이나 출처 불명의 데이터, 일자리 위협, 디지털 격차에 따른 사회적 수용성 부족을 해결하기 위한 정책이 필요하다. XAI와 같은 설명 가능한 AI 기술 확산과 시민 참여형 설계를 채택하고, AI 신뢰도 기준을 마련하며 AI의 '결과 설명창' 구현 표준화를 추진해야 한다. EU의 경우 AI ACT를 통해 분류된 위험 수

준에 따라 설명성 수준을 결정하는 등 XAI를 의무화하여 사용자에게 AI 판단 근거를 제공하고 있다.

고령층이나 저소득층의 접근성과 활용 역량을 강화하기 위한 디지털 배움터, 지역 교육 거점화를 확대할 필요가 있다. 세대별 수용성을 높이기 위해 중장년 대상 생활형 AI 교육 강화하고 지역 격차를 줄일 수 있는 지역 기반 AI 교육 인프라를 확보해야 한다. 영국은 2025년 1월, AI 그로스 존(AI Growth Zones) 프로그램을 통해, 지역별 'AI 성장 구역'을 지정하고 컴퓨팅 자원과 교육 중심지를 설치했다.

미래 교육을 위해서는 노동 변혁과 AI 에이전트 보편화에 적합한 핵심 역량 탐색과 교육 방향을 설정하는 작업이 필요하다. 또한 AI 에이전트 시대의 핵심 역량을 탐색하고 AI-시스템 연계형 교육 프로그램을 개발해야 한다. 특히, 직무 요구 역량 변화에 따라 직종별 산업 현장 실무중심형 AI 활용·실습 커리큘럼의 개발이 필요하다. 초·중등 AI 리터러시 강화형 교과 개편과 AI 교원 양성 제도도 도입되어야 하며 '(가칭)AI와 함께하는 협력 학습' 같은 교과 과정이나 AI 교사 인증제도 추진할 필요가 있다.

– 안성원(소프트웨어정책연구소 실장)

10. 사회적 수용성과 미래 교육
: AI와 함께 배우는 인간의 미래

AI 에이전트 기술에 대한 사회적 수용성이 비교적 높음에도 불구하고 '탈숙련' 문제와 사용자 주제성 문제 등이 해결해야 할 과제로 남아 있다. AI 에이전트의 편리함에 익숙해지고 과의존하게 되면서 개인적, 사회적 수준의 '탈숙련'이 발생할 수 있어서 이에 대한 대비가 필요하다. 개인적 수준에서 원래 할 수 있었던 인지적, 사회적 작업을 하기 어려워하거나 귀찮게 느끼게 되고 결국에는 할 수 없는 상황에 이르게 될 위험이 있다. 특정 전문 분야의 전문성을 배울 기회나 동기가 사라져서 궁극적으로는 사회 전체의 전문 역량이 '탈숙련'될 위험이 생긴다.

AI 에이전트 활용이 일상화되면서 자신의 삶에 대한 주체적 판단과 자신의 정체성에 대한 혼란 등이 발생할 수 있어서 이에 대한 대비가 필요하다. AI 에이전트 활용이 늘어나면서 스스로가 독립적으로 할 수 있는 일이 줄어든다는 위기감에 인지적, 정서적 무기력감에 빠질 위험도 있다. 자기효능감이 떨어지고 자신의 정체성에 혼란이 올 수도 있다.

이에 탈숙련 문제의 심각성을 인식하고 핵심 역량을 재규정할 필요가 있다. AI 활용도가 높아질수록 탈숙련의 위험이 높아진다. 현재 AI는 학술적, 교육적, 산업적 맥락에서 폭넓게 활용되고 있다. 교사나 교수 상당수는 AI를 과도하게 사용하면 학생들의 자율적 학습 능력이 훼손될 수 있다고 우려했다. 인간 두뇌의 신경 가소성(neuro plasticity)를 고려할 때, AI에게 사람이 통상적으로 수행하던 지적 작업을 습관적으로 외주화하면 종

합, 분석, 평가, 창의와 같은 고도의 지적 능력은 감퇴할 것이라 예상할 수 있다. 이는 이미 AI 등장 이전의 디지털 기술을 과도하게 활용함으로써 나타난 부작용이다. 일부 저자들은 이에 대응하기 위해 디지털 기술을 최소한 사용할 것을 권고하기도 했다.

AI의 활용 능력에 따른 'AI 격차(AI Divide)'가 예상되고 이에 대한 대응책이 모색되는 상황을 고려할 때, AI 활용을 막는 방식으로 '탈숙련' 문제에 대응하는 것은 바람직하지 않다고 판단된다. 그보다는 AI와 성공적으로 협업하면서도 특정 분야의 전문가로서 '역량'을 유지할 수 있는 방안을 찾는 것이 필요하다. 이를 위해서는 전문 영역별로 AI에게 어떤 경우에도 위임하지 않고 인간이 수행해야 하는 필수 역량, 필요에 따라 AI에게 위임하더라도 원칙적으로 필요한 상황에서는 인간 전문가가 수행할 수 있는 핵심 역량, 대부분 상황에서 AI에게 위임하더라도 큰 문제를 발생하지 않는 비필수 역량을 나누어 규정할 필요가 있다. 예를 들어, 의사에게 환자와 대화를 나누는 역량은 필수 역량으로, AI 도움 없이 환자를 진단하거나 치료할 수 있는 역량은 핵심 역량으로 지정하여 교육이나 전문의 훈련 과정에서 이들을 강조할 필요가 있다. 이는 전문가들이 AI 에이전트를 일상적으로 사용하는 상황에서도 체계적으로 유지되어야 하는 역량기 때문이다.

AI 에이전트를 과도하게 사용함에 따라 자율성에 위협을 받을 수도 있다. 어린 시절부터 AI 에이전트에 익숙한 AI 네이티브 세대는 AI 에이전트 없이 스스로 결정하고 행동하는 상황을 어색해할 수 있다. 이런 어색함이 심해지면 스스로 주체적으로 판단하거나 결정하고 행동하는 일 자체를 어려워하고 무엇이든 AI 에이전트에게 물어보고 결정하는 주체성 장애로 이어질 수 있어 이에 대한 대비가 필요하다.

또 이에 따라 정체성을 위협받을 수도 있다. AI 네이티브 세대는 자신의 개성을 발전시킬 기회를 갖지 못한 채 AI 에이전트가 추천하는 내용에 의해 자아가 과도하게 영향받을 위험이 있다. 그 결과 자신이 누구이고 어떤 가치를 추구하는지 등에 대한 분명한 정체성을 형성하지 못하고 AI 에이전트가 추천하거나 제시하는 내용을 맹목적으로 따르는 병리적 상태로 이어질 위험이 있다. 이에 대한 대응책도 필요하다.

사회적 수용성과 미래 교육 관련 기술의 긍정적 영향

AI 에이전트와의 효율적인 협업을 통해 인간 능력을 강화할 수 있다. 전문성을 충분히 갖춘 인간이 AI 에이전트와 효율적으로 협업하면 생산성을 높일 수 있다. 특히 전문가들이 AI 에이전트를 활용하면 '환각'을 줄일 수 있다. 또한 이미 축적된 전문가의 지식을 활용하여 보다 심도 깊은 분석을 진행할 수 있으므로 특정 주제에 대해 AI 에이전트의 도움을 받아 인간 사용자가 보다 효과적으로 통찰력을 발휘할 수 있다.

인간 전문가가 전문성을 획득하는 과정에 AI 에이전트가 도움을 줄 수도 있다. AI 에이전트가 전문가를 양성하는 과정에서 활용될 수도 있는데, 이를 위해서는 AI 에이전트 개발이 인간에게 '편리함'을 주는 방향이 아니라 인간 능력을 '계발'하는 방향으로 이루어진다는 전제가 만족되어야 한다. AI 에이전트가 인간 사용자에게 질문을 던지고 답을 찾거나 보다 심도 있는 분석을 하도록 돕는 방식의 기술 설계가 필요하다.

적절하게 설계되고 훈련된 AI 에이전트는 AI 네이티브 세대가 자신의 정체성을 형성하고 주체적으로 판단하고 결정하는 연습을 하는 과정에서 도움을 줄 수 있다. 그런데 이는 사용자에게 AI 에이전트에 대한 과의존을

경고하고 AI 에이전트가 제시하는 정보나 견해와 독립적으로 자신의 생각을 제시하고 이를 논증하도록 요구하는 AI 에이전트가 개발되고 활용된다는 전제 아래 진행되어야 한다. AI 에이전트의 개발과 활용이 현재처럼 지나친 상업적 이해 관계에 의해서만 결정되는 것이 아니라 사용자의 주체적 결정을 돕는 방식으로 개발되고 활용될 때만 기대되는 효과이다.

적절하게 설계되고 훈련된 AI 에이전트는 AI 네이티브 세대의 정체성 형성 과정에 바람직한 방식으로 도움을 줄 수 있다. 그런데 이를 위해서는 AI 에이전트가 사용자에게 통계적으로 전형적인 견해나 상황만을 전달하기보다는 사용자가 자신이 추구하는 가치를 선택하고 이와 정합적인 방식으로 자신의 생각과 행동을 결정해 나가도록 돕는 방식으로 개발되고 활용된다는 전제가 만족되어야 한다. AI 에이전트가 지나치게 '개인화'되어서 사용자에게 '아첨'만 하지 않고, 사용자가 자신의 정체성을 비판적 방식으로 형성해가는 데 방해를 하지 않는 방식으로 개발되고 활용될 때만 기대되는 효과이다.

사회적 수용성과 미래 교육 관련 기술의 부정적 영향

AI 에이전트의 편리함에 인간 사용자가 과의존하게 될 위험이 있다. AI 에이전트가 '알아서 척척' 해주는 서비스 품질이 향상됨에 따라 스스로 복잡한 생각을 하길 싫어하게 될 위험이다. AI 에이전트의 편리함에 익숙해져서 전문가로 성장할 기회를 스스로 포기하거나 힘들어 하게 될 수도 있다. AI 에이전트 활용이 늘어나면서 인간 사용자의 인지 능력이 감퇴할 우려도 있다. 원래는 인간 사용자가 할 수 있었던 인지적, 사회적 작업을 AI 에이전트 활용이 늘어나면서 점점 더 하기 어려워하거나 하기 귀찮아할 수도 있다. 사회 전반적으로 AI 에이전트에 대한 과의존이 심각해지면 정전 등의 기술적 장애로 AI 에이전트 활용이 불가능해질 때 대형 재난이 발생할 위험도 있다.

AI 에이전트 사용에 과몰입 혹은 중독되면 AI 에이전트 도움 없이 스스로 판단하거나 행동하기 어려워하는 상태에 이를 우려가 있다. 그 결과 특히 타인에 대한 평가나 판단을 스스로 자율적으로 수행하지 못하고 다른 사람의 평가나 판단을 AI 에이전트를 통해 수집하여 이를 자신의 판단으로 대체하게 될 수도 있다.

아주 어린 시기부터 AI 에이전트를 사용한 AI 네이티브는 AI 에이전트의 전형화된 조언이나 판단을 내재화하여 자신의 정체성을 형성할 위험에 처할 수 있다. 그 결과 자신이 추구하는 가치가 어떤 이유에서, 어떤 근거를 갖고 있는지에 대한 고민 없이 AI 에이전트가 제공하는 대중적 경향성을 무비판적으로 자신의 정체성으로 받아들일 가능성이 있다. 이 경우 복잡한 가치가 충돌하는 상황에서 정체성 혼란이 오고 중요한 사회적 결정에 대해 일관된 입장을 취하지 못하게 될 우려가 있다.

– 이상욱(한양대학교 교수)

AI 생성 이미지

11. 이용자 특성 평가
: 모든 사람이 AI를 같은 방식으로 받아들이지 않는다

AI 에이전트는 기술적 성능에 더해 사용자가 가진 사회적·심리적 특성에 의해 수용 여부가 크게 달라지는 대표적 기술이다. 사용자는 AI 에이전트를 기술이 아닌 사람처럼 대하는 경향이 반복적으로 보고되고 있다. 따라서 이용자 특성과 에이전트 특성이 상호 작용하여 복잡한 반응 양상을 보일 수 있다. 젠더, 연령, 사회적 지위, 장애, 문화 등 이용자 특성에 의해 접근성·정확도·위험이 비대칭으로 나타나며, 데이터 편향이 사회적 불평등을 재생산할 수 있다.

성별은 AI 에이전트 이용 과정에서 중요한 역할을 하는 요인이다. 시리, 알렉사, 구글 어시스턴트 등 여성 음성 어시스턴트가 성 역할 고정관념을 강화할 수 있으므로 개발자들은 이에 대한 고려가 필요한 것으로 보고되었다. AI 에이전트 이용 환경에 따라 성별의 영향이 다르게 나타났다. 기능 중심 환경에서는 남성 AI 에이전트가 더 높은 신뢰 수준을 이끌어냈고, 경험 중심 환경에서는 여성 AI 에이전트가 더 강한 안정된 몰입감을 제공했다. 협력적 상호 작용 환경에서 사람들은 여성으로 라벨된 AI 에이전트를 더 자주 이용하려 했고, 남성 AI 에이전트는 불신하는 경향을 보였다.

이용자 특성 평가 관련 기술의 긍정적 영향

고령자나 기술에 익숙하지 않은 사용자들에게 AI 보조자는 음성 인터페이스 또는 간단한 명령 체계를 통해 접근 장벽을 낮추고, 독립적 삶(in-

dependent living) 유지와 사회적 고립 감소에 기여하고 있다.

외로움(loneliness)이 크거나 사회적 활동이 제한된 사용자에게 AI 에이전트는 대화 상대 또는 동반자(companion)의 역할을 함으로써 정서적 안정, 우울감(depression)과 스트레스 지표 감소 가능성이 있다. 에이전트가 감성(emotional) 대응 능력 또는 사용자 취향(personality/affective traits)을 고려할 때, 사용자 몰입(engagement)과 만족도가 유의미하게 증가됨을 볼 수 있다. 사용자의 과거 경험, 선호, 상황(context)을 반영하여 반응이나 추천을 조정함으로써, 사용자는 더 덜 반복적이고 더 유용한 상호작용을 경험하게 된다. 이로 인해 시간 절약, 결정 비용(decision cost) 감소 가능성이 있다.

AI 에이전트의 생산성 향상 효과는 이용자 직업, 연령, 디지털 숙련도, 조직 내 지위에 따라 서로 다른 방식으로 나타난다. 지식노동자는 문서 작성·회의 요약에서, 현장·서비스직은 일정·재고 관리에서 생산성 효과가 크다. 젊은 세대는 멀티태스킹 지원, 고령자는 음성 기반 단순화로 효율성을 체감하고 디지털 숙련도가 높은 이용자는 업무 위임을, 낮은 이용자는 기초적 반복 작업 지원을 주로 활용하는 것으로 나타난다. 관리자와 실무자는 각각 의사 결정 지원, 반복 업무 자동화에서 효과를 얻고 있다.

이용자 특성 평가 관련 기술의 부정적 영향

여성 목소리를 기본으로 설정하는 것은 '보조자' 이미지와 맞물려 성역할 고정관념을 강화한다. 비서 역할을 하는 에이전트에게 여성 목소리를 기본 설정으로 제공하는 것이 일반적이다. 성별에 따라 신뢰·몰입·착취 경향이 다르게 나타나며, 장기간 누적될 경우 사회문화적 성별 편향을 재생산할 수 있다.

고연령층, 종교적 신념 집단, 통제와 자율성을 중시하는 집단은 의인화된 AI 에이전트에 대해 거부감·불안을 가질 수 있다. 프라이버시 민감 집단은 AI 에이전트와의 상호 작용을 통제 상실로 인식해 사용을 회피할 가능성이 크다.

VIP·프리미엄 사용자에게는 정교한 서비스를 제공하고, 일반 사용자는 단순화된 서비스를 받는 구조가 형성될 경우 서비스 품질 불균형이 심화될 수 있다. 가령 VIP 고객이나 특정 계층은 더 정교한 응답을, 일반 사용자는 단순화된 서비스만 제공받는 구조가 생길 수 있다. 기술 권한이 자본과 권력을 가진 집단에 집중되며, 불이익에 대한 책임 소재 불명확성이 취약 계층에서 특히 문제화될 수 있다.

AI 에이전트는 대규모 언어 데이터와 인간 피드백(RLHF)을 통해 학습되는데, 이 데이터가 특정 문화·언어권에 편중될 경우 다른 문화적 맥락에서는 오해나 불일치가 발생할 수 있다. 직설적 대화 양식을 선호하는 서구 문화와 간접적 화법·맥락적 소통을 중시하는 동아시아 문화 간의 차이는, AI 에이전트가 협상·의사소통 상황에서 적절히 대응하지 못하게 만들어 사용자 경험을 저해할 수 있다.

기존 연구에서는 AI 언어 모델이 영미권 데이터에 과대 표집되어 동남아·아프리카 사용자에게 문화적 불일치를 일으킨 사례를 보고했다. 이는 글로벌 서비스 환경에서 경제적 손실이나 협상 실패·국제 불신으로 이어질 수 있다. 문화적 편향은 단순한 불편을 넘어 사회적 불평등 재생산으로 이어질 수 있으며, 특히 다문화·이민 사회에서는 특정 집단을 주변화시키는 위험을 안고 있다. 또한 문화적 불일치가 국제 협상이나 경제적 활동 과정에서 발생할 경우, 이용자에게 실질적 피해가 전가될 수 있다.

정책 수요와 대안

사용자 특성을 반영한 공정·포용적 AI 에이전트를 개발하기 위해 성별, 연령, 사회적 지위, 장애, 문화적 차이를 포괄하는 다층적 사용자 프로파일링(multidimensional user profiling)을 개발·테스트 과정에 반영해야 한다. 단순 기능 검증을 넘어, 사용자 경험 기반 평가(user-centered evaluation)를 제도화하여, 특정 집단의 배제·차별 가능성을 사전에 차단할 필요가 있다.

기존 "Bias Impact Assessment(BIA)" 개념을 확대하여, Algorithmic Impact Assessment(AIA) 및 AI Risk Assessment Framework를 도입하고 적용하여 알고리즘 영향평가와 편향 검증을 제도화해야 한다. 평가 항목에는 성별·연령·소득 수준뿐 아니라 문화적 특성·지역적 맥락을 반영하여 결과의 문화적 정합성(cultural alignment)을 보장해야 한다.

OECD(2019), 유네스코(2021), AI Act(2024)에서 제시한 문화적 다양성·포용성 원칙을 반영하여, AI 에이전트의 문화 공정성 검증(cultural fairness audit/ cross-cultural AI audit) 제도를 운영할 필요가 있다. 한국적 맥락에서 간접적 소통 양식, 맥락 중심 의사 표현 등 동양권 커뮤니케이션 특성을 반영한 '문화 친화형 AI 에이전트 가이드라인'을 수립하고 나아가 사용자가 성별·문화·연령대 등 에이전트 특성을 선택(customizable features)할 수 있는 기능을 제공하여, 이용자 주도권을 강화해야 한다.

개발자·서비스 제공자·정부 등 에이전트 이용 과정에서 발생하는 문화적 불일치·불평등에 대한 책임 주체의 역할 분담을 명확히 규정해야 한다. 국제적으로 통용되는 책임성(accountability)·투명성(transparency) 원칙에 기반한 제도적 거버넌스 구축이 필요하다.

– 이재신(중앙대학교 교수)

12. 시민사회와 민주주의 : AI 에이전트의 참여와 침투

AI 기술, 특히 에이전트 기반 AI가 정치·사회적 의사 결정, 정보 유통, 공공 담론 형성 등에 어떻게 영향을 미치는지를 분석하는 영역인데 크게 '시민 참여와 정치'와 '여론과 정보 다양성'이란 두 가지 이슈로 살펴볼 수 있다.

AI 에이전트는 단순한 도구를 넘어서 정치적 매개자, 정보 필터, 사회적 소통의 촉진자 또는 장애물로 작용할 수 있다. 디지털 사회에서 AI 에이전트는 '제4의 시민'처럼 작동할 가능성도 있다. 자유, 평등, 참여 등 민주주의 3대 원칙이 AI에 의해 강화될 수도, 약화될 수도 있다.

AI 생성 이미지

AI 에이전트를 통한 시민 참여 플랫폼은 확산 중으로, 공공 정책 형성 과정에서 AI가 시민의 의견을 매개하고 요약·전달하는 역할을 하게 된다. 그 사례로 국민신문고, 국민권익위원회 중심으로 AI 기반 민원 분석과 자동 분류·추천 기능이 도입되기 시작한 것을 들 수 있다.

AI 에이전트는 행정, 복지, 보안 등 다양한 공공 분야에서 정책 조언자 또는 대리 결정자로 활용될 수 있다. 하지만 AI 에이전트 기반 정책은 정책 과정에 시민의 의견과 감시 기능을 제도적으로 확보해야 민주적 정당성을 얻을 수 있다.

독일 바덴뷔르템베르크주(州)의 시민 배심회 'AI 앤드 프리덤(AI and Freedom)'을 사례로 들 수 있다. 튀빙엔대학교가 주최하고 폭스바겐재단이 후원하여 다양한 지역과 계층에서 무작위로 선정된 시민 40명이 참여하는데 공공자금으로 진행되는 AI 연구 개발 과정에 사회적 가치와 시민의 관점을 반영할 목적으로 시행되는 프로젝트이다. 2025년 3월 공동 권고안(9개 권고안)을 과학장관에게 공식 제출. AI 기술에 대한 시민의 의견을 구조화된 정책 제안으로 전환한 대표적 사례이다. AI의 윤리적 판단 기준에 관한 사회적 합의, 대의제 한계 보완, 기술 불신 해소, 공공 감시 기능 강화를 위해 시민 의견 반영이 필요하다.

AI 에이전트는 여론의 기저를 형성하는 정보 생태계의 핵심 행위자로 떠오르고 있다. AI 에이전트는 사용자의 관심사, 검색 이력, 발화 습관 등을 학습해 개인 맞춤형 정보를 제공하는 역할을 한다. AI 에이전트가 정보 필터링과 요약, 추천을 통해 여론 형성에 직간접적 영향력을 행사한다. 기존 언론이나 미디어 플랫폼 중심에서 AI 에이전트 중심의 여론 형성으로 바뀌고 있다. 이는 시민이 AI 추천과 요약의 한계를 인지하고 능동적으로 대처할 수 있어야 가능하다.

AI 에이전트 시대에 시민이 정보 환경을 이해하고 참여할 수 있는 권리와 역량 갖춰야 한다. 또 AI 에이전트가 불러올 변화에 대해 시민이 충분히 이해하고 참여할 수 있도록, 다양한 의견과 사례 전달하고 AI 기술 관련 잘못된 정보나 과장된 기대를 걸러내고, 과학적·객관적 근거에 기반한 사실을 전달하는 것도 필요하다. 이로써 AI 에이전트를 둘러싼 사실 기반의 과학 커뮤니케이션과 저널리즘이 중요해지고 있다.

시민사회와 민주주의 관련 기술의 긍정적 영향

AI를 활용한 정책 모니터링, 청원 분석, 공공데이터 요약 등의 기능을 통해 오히려 시민 참여를 확장할 수 있다는 기대도 있다. AI 에이전트는 시민이 감시자 역할을 할 수 있도록 도와주어, '하향식 행정'이 아니라 '상향식 의사 결정' 구조로 전환되는 계기를 마련할 수 있다. AI 기반 참여는 언제 어디서나 가능해 물리적 회의나 설문 조사보다 훨씬 유연하고 지속적인 참여 구조를 만들어내 시간과 장소의 제약을 극복하게 해준다. 그래서 지방 분권이나 마을, 학교, 단체 등 소규모 단위 정치에서도 활용 가능성이 크다.

AI 에이전트는 대규모 의견을 수렴하고, 논의 구조를 시각화하거나 요약함으로써 시민 의견이 정책에 반영되는 경로를 효율적이고 투명하게 만들어준다. 캐나다 Algorithmic Impact Assessment(AIA)의 사례를 들 수 있는데 이는 캐나다 연방정부가 2019년부터 시행한 디지털 정부 전략의 일환으로 시행되었다. 자동화된 의사 결정 시스템을 정부가 도입하거나 운영할 때, 그 시스템이 미칠 수 있는 영향의 수준을 사전에 평가하도록 요구하는 것이다. 시민의 의견을 사전 수렴하고, 알고리즘의 위험 수준에 따라 공론화 여부를 결정한다.

AI 에이전트는 복잡한 정책 내용을 자동으로 요약하고 번역·설명해주며, 다양한 언어, 문해 수준을 고려하여 맞춤형 정보를 전달할 수 있다. 이로써 시민들이 정책과 정치에 더 잘 접근하고 이해하도록 도우며, 숙의민주주의 기반을 다져준다. 여기에는 에스토니아의 사례를 들 수 있다. 전자정부 시스템에서 AI 챗봇이 행정 정보와 정책 절차를 시민에게 쉽게 설명한다. 대표적 예로 통합 챗봇 네트워크 뷰로크라트(Bürokratt), 위기 대응 전용 AI 챗봇 수베(Suve)를 들 수 있다.

AI 에이전트는 사회 각계각층의 다양한 의견을 자동 수집하고 분석해, 공론장에서 쉽게 지나쳤던 소외자, 소외 지역, 청소년, 고령층 등의 의견을 드러내줄 수 있다. 다언론 시대의 혼잡 속에서도 정제된 다수의 관점을 요약하거나 시각화해줄 수 있다. 또 AI 에이전트는 개인의 언어 수준, 시각·청각 장애, 디지털 문해력 등에 맞춰 맞춤형 정보를 제공함으로써 정보 접근성을 향상시켜준다.

시민사회와 민주주의 관련 기술의 부정적 영향

AI 에이전트가 시민 의견을 대리하거나 요약함으로써, 실제 민의를 기계적 해석에 의존하는 문제가 발생하게 된다. 자동화된 참여는 실질적 참여의 왜곡, 정치적 개입, 의견 반영의 비대칭성을 초래할 수 있다. 특히 권위주의 체제에서는 AI가 감시, 선별, 검열의 수단으로 악용될 가능성도 크다. 또 정보 격차가 심화되고 소외 집단이 배제될 가능성도 있다. 디지털 접근성이 낮은 집단이나 사회경제적 약자가 AI 기반 정책 참여 플랫폼에서 배제될 위험이 있는 것이다.

AI 모델이 어떤 데이터를 학습했고, 어떤 기준으로 판단이나 추천을 했는지에 대한 설명 가능성이 없다면, 시민은 AI를 통한 정책 결정 과정 자체를 신뢰하지 않게 되고, 공공 정책의 정당성이 흔들릴 수 있다. AI 에이전트가 운영하는 콘텐츠 추천 알고리즘은 사용자의 클릭, 시청 이력 등 개인화된 데이터를 기반으로 작동하므로, 사용자가 기존에 선호하던 시각만 접하는 '필터 버블(filter bubble)'을 심화시킬 가능성도 있다. 다양한 의견을 접할 기회를 박탈함으로써 정치적 양극화, 소수 의견의 배제, 사회적 단절 촉진할 우려가 있다. 특히 사회적 소수자나 지역 기반 소통에서 정보 접근권의 불균형이 악화될 수 있다.

극소수 기업에 의한 AI 기술의 독점 가능성이 있으며, 다양성 파괴의 우려가 존재한다. 기술에 대한 공동 통제가 미비한 상황에서 AI를 소유한 기업과 정권이 결합할 경우 정보, 정책, 여론을 통제하는 초집중 권력이 형성될 수 있는데 이는 헌법적 권리 약화, 의회와 시민 참여의 무력화, 소수 의견의 소거로 이어질 수 있다.

출처가 불명확하거나 검증되지 않은 정보, 왜곡된 사실, 과장된 기대가 포함된 콘텐츠가 무분별하게 유통될 수 있다. 신뢰할 수 있는 정보에 접근할 시민의 능력을 떨어뜨리고, 허위 정보 또는 과장된 기대가 정책 논의 왜곡, 여론 조작, 시민 오판을 유도할 수 있다.

정책 수요와 대안

AI 에이전트 기반 시만 참여 플랫폼에 대한 제도 정비가 필요하다. 공공 정책 참여용 AI 에이전트가 도입될 가능성에 대비해야 하고 시민 대표성, 중립성, 투명성을 확보할 방안이 필요하다. 이를 위해서는 디지털 시민 토론 에이전트 윤리 가이드라인을 제정해야 한다. 공공기관 주관 AI 토론 플랫폼을 구축할 때 설명 가능성, 기록 보존, 토론 다양성 등의 보장을 의무화해야 한다. 그 사례로 덴마크 디지털 시민회의(Digital Citizens'

Assembly)를 들 수 있다. 여기서는 국민이 AI 기반 질의 응답 시스템을 활용해 환경·복지 정책 의견 수렴에 참여하고 있다.

설명성·공정성 기준을 통과할 때 인증해주는 공공 AI 에이전트 중립성 인증제를 도입하는 것도 고려할 만하다. 현행 「전자정부법 시행령」 제57조 제2항: 공공 SW의 품질보증 및 인증기준 마련 규정에 신뢰성, 기능성 등 평가 기준을 포함하는 것이다. 이런 사례로 UK Government Digital Standards(2021)를 들 수 있다. 특히 Algorithmic Transparency Recording Standard(ATRS)는 공공기관 AI 알고리즘에 설명 가능성·공정성 기준 적용을 권고하고 있다.

선거와 정치적 발언 영역에서의 AI 에이전트 활동 규제도 필요하다. AI 에이전트가 특정 정당·후보를 지지하거나 허위 정보를 전파할 수 있기 때문이다. 그런데 '정치적 표현'과 '조작' 사이 경계가 불분명하다는 문제가 있다. 이를 해결하기 위해 AI 에이전트의 정치 콘텐츠 생성·확산 제한 조항을 신설하는 것을 검토해야 한다. 선거 기간 중 정당이나 제3자가 사용하는 AI 에이전트의 정보 생성·확산에 대한 투명성과 출처 표시를 의무화하도록 하는 것도 방법이다. 그 사례로 EU 디지털서비스법(Digital Services Act, 2022)을 들 수 있는데 선거 기간 중 정치 광고에 AI 활용 시 자동화 여부와 주체 공개할 의무를 규정하고 있다. 현행 「공직선거법」 제82조의6: 선거 정보 매체의 출처, 실명, 허위 정보 제한 조항이 있다.

선거 기간 중 AI 자동 콘텐츠 '사전 등록제'를 운영하고 자동 콘텐츠 시스템은 선관위에 등록·감시 대상 포함하는 방법도 있다. 현행 「공직선거법」 제110조: 인터넷 매체 선거 광고에 대한 사전 신고 및 심의 절차 규정에 준하는 것이다. 이런 사례로는 미국 연방선거위원회(Federal Election Commission, FEC)를 들 수 있다. 여기서는 2023년부터 AI 사용 캠페인 콘텐츠에 대한 사전 공개와 규제 검토를 착수했다.

AI 에이전트가 자동화된 알고리즘으로 편향된 여론을 형성할 우려가 있고 개인화된 정보 추천이 확증 편향, 필터 버블(filter bubble) 현상을 심화할 수 있어 뉴스 편향, 정치 양극화를 유도할 수도 있다.

이의 해결을 위해서 AI 에이전트의 정보 추천 알고리즘에 '다양성 기준'을 의무화할 필요가 있다. 여기 정치적 성향·출처 다양성을 일정 비율을 포함하도록 가이드를 제정하는 것이다. 그 사례로 독일 NetzDG법 개정안(2021)을 들 수 있다. 여기서는 알고리즘 기반 콘텐츠 필터링의 투명성을 요구하고 있다. 현행 「방송통신발전기본법」 제17조, 「정보통신망법」 제22조의5에 정보 추천 시 투명성·이용자 통제권 부여 근거가 있다.

공공 알고리즘 감시를 통한 '정보 다양성 지수' 공개 제도를 도입할 수도 있다. 플랫폼 기업이 사용하는 에이전트 추천 알고리즘에 대해 '다양성·균형성' 관련 감사 지표를 정기적으로 공시해야 한다. 프랑스 디지털 공화국법(Digital Republic Law)을 사례로 들 수 있는데 행정기관이 알고리즘을 기반으로 한 개별 행정 결정을 내릴 경우, 해당 알고리즘의 처리 방식, 주요 파라미터, 데이터 출처 등을 설명해야 하며, 시민은 이를 열람하고 이의를 제기할 권리를 가지고 있다. 이는 단순 결정뿐만 아니라 알고리즘 기반의 의사 결정 보조 시스템에도 명시적으로 적용되는 법률적 권리 구조이다.

현행 「방송법」 제86조의 2: 방송의 공정성, 균형성, 다양성 확보 규정으로 다음과 같은 정책적 활용 가능성이 있다.

1) AI 기반의 콘텐츠 추천 플랫폼이 여론 형성에 미칠 수 있는 영향에 대비해 추천 알고리즘의 다양성 기준 적용 가능성 제고
2) AI 에이전트가 생성·추천하는 정치·사회 콘텐츠에 대해서도 이 규

범 내에서 균형성·투명성 확보 조치로 해석 가능

AI 에이전트의 추천 알고리즘에 기존 언론이 휘둘리는 위험도 있다. AI 에이전트가 만든 가짜 뉴스가 쉽게 퍼져 언론의 역할이 약화되고 있다. 이를 해결하기 위해서 '공인 뉴스 콘텐츠'에 대한 알고리즘상 공공 우선 배치 규정을 마련할 필요가 있다. 기후, 건강, 선서 등 공공 의제 관련 뉴스 콘텐츠는 알고리즘상 우선 노출을 의무화하고 일정 비율을 확보해주는 것이다. 캐나다 온라인 뉴스법(2023)를 사례로 들 수 있는데 여기서는 디지털 플랫폼이 뉴스를 제공할 때 외국 콘텐츠보다 캐나다 뉴스를 우선 배치하도록 유도하고 있다. OECD 보고서/로이터연구소(Reuters Institute) 보고서(2024~2025)에서는 디지털 플랫폼과 알고리즘이 미디어 다양성을 왜곡하고, 사용자 특성에 따라 정보 접근 격차를 불러오는 현상을 경고하고 있다.

전통 언론과 AI/플랫폼 기업 간 공동 팩트 체크 네트워크를 구축하는 방법도 있다. AI 에이전트가 생성하거나 추천하는 콘텐츠의 신뢰성을 검증하기 위해 언론사, 플랫폼, 공공기관이 공동으로 참여하는 팩트 체크 구조를 마련하는 것이다. 대표적 사례로 독일 비영리 탐사 보도 미디어기관 코렉티브(CORRECTIV)를 들 수 있다. 여기서는 언론사, 플랫폼 연합 팩트 체크 기구를 운용하고 있다. 한국 팩트체크넷(KPF)은 2020년 한국언론진흥재단이 설립한 시민 참여 기반 팩트 체크 플랫폼인데 향후 AI 추천 시스템과 연동할 필요가 있다.

– 이충환(동아에스앤씨 이사)

13. 윤리 가치

AI 윤리는 AI가 가져야 하는 기본적인 동작 철학이나 생각에서 우선순위에 해당되는 영역으로, 사람 먼저, 안전 먼저, 사회적 규범 안에서 기술을 개발하는 것을 의미한다. AI 윤리에는 사회적 윤리, 집단적 윤리, 개인적 윤리 기준이 있으며, 이들은 상호 간에 충돌하고 지속적으로 성장하고 있다.

사회적 윤리는 사회의 구성원들이 동의하는 윤리적 사항으로 통상적으로 사회 구성원이 모두 공유하는 윤리로 본다. 특히 사회적 규범이나 전통적 가치관에 따른 윤리 사항을 말하며, 사람 먼저, 안전 먼저와 같이 절대

AI 생성 이미지

적 윤리가 있고, 가치 기준에 따라 변화하는 윤리가 있다.

집단적 윤리는 개인이 속한 집단의 윤리를 말하며, 통상적으로 직업군에 의한 윤리 의식에 해당된다. 의사라면 생명을 최우선으로 하고, 변호사라면 변호인의 이익을 최우선시할 것이다. 따라서 윤리적 관점에서 집단적 윤리는 매우 한정적으로 편향될 가능성이 있다. 개인적 윤리는 개개인이 가지고 있는 윤리적 사고로 사람마다 다른 양상을 가지고 있다. 대부분은 사회적 윤리에 포괄적으로 들어가나 개인적으로 금전적 가치를 우선시하거나 사람의 관계를 우선시하는 등의 차이가 발생할 수 있다.

윤리 의식 또는 사회적 윤리는 비교적 변하지 않지만 시대적 관점에서는 다소 변화가 있을 수 있다. 특히 집단의 윤리나 개인의 윤리는 계속 변화하는 것으로 이를 반영한 사회 윤리 또는 윤리 의식을 만드는 것이 필요하다. 사회는 경제 상황, 지정학적 위험 등에 따라서 변화한다. 한반도는 전쟁 위험이 크기 때문에 정치적 갈등이나 보수, 진보의 치열한 싸움이 있으며, 사회적 안정에 따라서 충분히 변화하고 발전하게 된다.

집단의 경우, 내가 속하는 집단이 바뀌는 경우도 있지만 집단의 이익이나 생각에 변화가 생기길 수도 있다. 과거 핵기술을 연구하는 연구자는 사명감이 강했다면, 지금은 핵기술을 의료 등 민간에 활용하는 기술에 집중하면서 그 가치나 활용하는 방법에서 차이를 보이고 있다.

개인의 변화는 사회의 관계성이나 사회 구성원들의 상황에 따라 달라지는데 내가 경제적 어려움이 있는 경우 나의 생명을 위해서 경제 문제가 강조되는 것은 당연하다. 따라서 다양한 개인의 상황에 따라 지속적으로 변화하는 가치라 할 수 있다.

다양한 윤리적 가치가 존재할 수 있으며, 이들은 끊임없이 변화하고 또한 충돌하고 있다. 따라서 AI 기술도 이러한 윤리적 변화를 끊임없이 배우고 따르는 것이 필요하다. 집단적 윤리는 보편적인 윤리보다는 집단의 특성을 반영하는 윤리이기 때문에 사회적 지탄을 받는 경우가 많다. AI가 이러한 집단적 이기심을 배우는 경우 사회적 혼란이 있을 수 있다. 따라서 집단적 윤리보다는 사회적 윤리가 우선시되어야 하며, 이러한 큰 틀에서의 윤리적 관점을 반영하는 다양한 연구가 필요하다.

윤리 관련 기술의 긍정적 영향

점차 AI의 윤리적 관점이 사회적 이슈로 대두될 것이며 이를 적극적으로 반영하는 방법에 대한 논의가 시작될 것이다. AI 활용에 있어 큰 걸림돌 중 하나가 윤리적 판단을 할 수 있는가이며, 이를 극복하기 위한 방법으로 사회적 가치 변화를 AI가 학습할 수 있음은 활용 측면에서 매우 긍정적이다.

사회적 가치를 평가하는 것은 매우 어렵지만 사회적 합의를 얻기 위해서는 가치 판단의 근거를 제시해야 한다. 변화하는 사회를 빠르게 반영하는 것과 기존의 사회적 합의를 반영하는 것으로 모두 중요하다. 변화하는 사회의 가치를 반영하고 기존 가치를 변화하는 방향에 맞춰 지속적으로 사회에 공유하는 활동은 사회 통합의 차원에서 긍정적인 활동이다.

개인의 입장에서 가치의 완성은 사회적 가치, 집단적 가치, 개인의 가치를 모두 충족할 때 이뤄지기 때문에 이들 간의 충돌을 해결하는 것은 사회적 가치 완성에 긍정적이다. 가치의 크기와 달리 개인 입장에서 개인의 가치, 집단의 가치가 더 크기 때문에 이를 적절히 융합할 수 있는 기술이 필요하다. 서로 대립하는 가치는 이를 해결하면서 사회적 융합을 이루는 기술로 성장할 수 있다.

사회적 가치에서 가치 간의 충돌이 있는 경우 사람도 적절히 판단할 수

없으나 이들 사이의 치열한 논쟁은 가치 판단의 성장을 이끌 것으로 보인다. 암묵적으로 사회적 가치를 강조하기보다는 가치 간의 충돌에 대해서 다양한 논쟁을 하고 사회적 합의를 이끌어내는 것이 갈등 해소에 도움이 된다. 사회적 문제가 발생하는 경우 한쪽의 일방적인 손해보다는 서로의 가치 판단을 공유하는 것을 통합을 위한 좋은 방법이다.

윤리 관련 기술의 부정적 영향

AI의 윤리적 판단에 따른 평가가 바뀔 수 있어 논쟁의 중심이 될 수 있다. 특히 사회적 평가에 변화가 생기는 시기에는 사회적 갈등과 혼란을 야기할 수 있다. 특정 집단의 윤리 의식에 문제가 발생하는 경우 이를 바로잡기 위한 많은 노력이 필요하고 사회적 문제성을 만들 수 있다.

윤리적 가치 충돌은 개인보다는 집단에 의해서 결정되고 힘이 압도적으로 큰 집단에 의해서 결정될 가능성이 있기 때문에 사회적 필요보다는 개인적 필요가 작용할 수 있다. 부의 차이는 다른 영역에서도 큰 차이를 만들 수 있다. 잠시 욕을 먹지만 사회적인 통념을 바꾸면 막대한 이익이 생기는 경우 강제적으로 윤리적 가치를 바꾸려 할 수 있다. 사회의 여론을 형성하는 과정에서 다수의 침묵은 일부에 의해서 주도될 수 있어 사회적 부작용이 될 수도 있다.

윤리적 가치는 사회적으로 큰 영향력을 끼친다. 그 다음으로 집단이나 개인의 가치가 중요하지만 개인 입장에서는 반대로 개인이 가장 중요하고, 다음으로 집단, 사회 순이다. 따라서 윤리적 가치를 판단하는 기준이 개인의 가치와 다를 수 있다. 길거리에서 돈을 주운 경우, 사회적으로는 주인에게 찾아주는 것이 맞다. 하지만 개인적 욕심에서는 유용하고 싶은 것이 당연하다. 따라서 이러한 윤리적 기준이 작용할 때 개인의 선택을 중심으로 사회적 합의가 필요하다. 윤리적 가치는 단순한 돈으로 판단할 수 없으므로, 사회적 가치를 발굴하지 않는다면 올바른 가치 판단을 할 수 없다.

정책 수요와 대안

AI 에이전트에서 활용되는 다양한 데이터 활용에 대한 적극적인 가이드라인 개발이 필요하며 특히 윤리 충돌을 해결하기 위한 방법을 개발해야 한다. 사회적 윤리 의식 변화를 반영하여 사회의 변화를 모니터링하고 AI가 이를 학습할 수 있는 기술도 개발해야 한다. 사회적으로 변화하는 윤리 가치를 표현하고 새롭게 정립되는 가치를 반영할 수 있는 연구도 개발되어야 하는데 그 기술은 지속적으로 변화하는 사회적 윤리 가치, 집단적 윤리 가치, 개인적 윤리 가치를 식별하고 정립하는 것이어야 한다. 사회적 윤리, 집단적 윤리, 개인적 윤리 사이의 충돌에 대한 다양한 해결 방법도 적극적으로 연구해야 한다.

– 박종열(서울과학기술대 교수)

AI 생성 이미지

14. 융합 및 혁신
: AI 에이전트가 만드는 새로운 협업 질서

생성형 AI가 '응답형 도구'에서 목표 지향·연속 행동이 가능한 AI 에이전트로 고도화되고 있다. 이는 외부 도구 호출, 다중 단계 계획, 다른 에이전트와의 협업까지 포함하고 있는데 더 이상 단일 기능 자동화 도구가 아니라, 문제 해결, 의사 결정, 사용자 맞춤형 지원 등 복합 기능을 수행하는 '디지털 동반자'로 진화하는 것이다.

〈AI 에이전트의 다섯 가지 주요 특징〉

특징	내용
인식 (Perception)	센서, 데이터 파일, 인터넷 등 다양한 데이터 수집 방법을 통해 주변 환경 인식 및 파악
처리(Brain)	입력 데이터 기반 주요 정보 추출 및 새로운 정보 학습을 통한 의사 결정
행동(Action)	도출된 결정, 해답 기반 텍스트 생성, API 호출, 물리적 동작 등 실제 수행
학습 및 적용 (Learning and Adaptation)	경험 기반 지속적 학습, 과거 데이터를 기반으로 미래 행동 개선
자율성 (Autonomy)	스스로 결정을 내리고 행동하며 인간의 개입 없이 목표를 달성하기 위한 작업 수행

로봇·IoT·디지털트윈·클라우드/엣지·RPA(Robotic Process Automation)와 결합 시 물리·디지털·제도의 경계가 약화되고 제조·검사·품질 관리 과정을 포함한 공장, 진단·청구·사후 관리가 필요한 병원, 온보딩·

〈AI에이전트 관련 주요 관련 법·규범〉

관련 법/제도	주요 내용
개인정보보호법	데이터 활용 규제 완화 및 명확화
정보통신망법	AI 자율 의사 결정 시 책임 주체 불명확
산업안전보건법	인간, 에이전트 협업 환경에서 안전 규제 미비
상법	AI 대리 행위 관련 명확성 부족
전자금융거래법	AI가 금융 의사 결정 및 거래 수행 가능성 근거 부족
국제규범(EU AI Act 등)	글로벌 규제 정합성 확보 필요(고위험 AI 분류 평가 체계 등)

심사·사기 탐지 등을 포함한 금융 등 엔드투엔드 통합이 가속되고 있다.

AI 에이전트 특성으로 인한 맞춤 서비스가 가능해짐에 따라, 공공 서비스, 의료, 복지 등 기존 시스템의 한계를 보완하고 고도화할 수 있는 잠재력이 크다. 이에 따라 과학 기술, 산업, 정책, 사회 전 분야에서 융합적 사고와 구조적 혁신을 병행할 필요가 증가하고 있다.

자율 에이전트의 행위 책임, 편향이나 안전성, 설명 가능성 확보가 사회적 수용성의 전제가 될 수 있다. 개인 정보·저작권·모델 훈련 데이터의 합법성, 데이터 이동권/동의 철회, 로그·감사 체계 확립 전제로 수요가 늘어날 수 있다.

융합 및 혁신 관련 기술의 긍정적 영향

ERP·CRM 등 다른 기종 시스템 사이를 툴 호출·API 어댑터·이벤트 버스로 연결하여 부서·기관 간 데이터/업무 흐름을 자연스럽게 연동시킬 수 있다. MCP, A2A 등 커넥터 방식 출현으로 LLM/에이전트가 외부 데이터·툴에 붙는 방식을 표준화해 한 번 구현으로 다수 시스템과 재사용이나 확장이 가능하다.

AI 멀티에이전트를 활용하면 '단일 기능'이 '연결형 번들'로 전환될 수 있다. 로봇과 에이전트가 만나면 검사-피킹-포장-출하를 하나의 협업 시나리오로 연계할 수 있고, 안전·품질 데이터를 상시 학습할 수 있다. 의료와 보험이 만나면 진단·청구·사후 관리를 전주기 자동화하여 누락이나 지연을 감소시키고 의료-보험 데이터 브릿지를 통해 의료비 투명성을 제고 높일 수 있다.

소재나 제조와 에이전트가 협업하면 실험 설계-시뮬-공정 레시피 최적화가 단일 워크플로로 통합되어 R&D-양산 간 간극을 축소할 수 있다. 오케스트레이션 기능으로 단순 병렬이 아닌, 목적 지향적인 워크플로우 구성하고 신뢰 가능한 에이전트 생태계를 구성하는 것도 가능해진다.

구글·오픈AI 등 글로벌 기업이 AI 에이전트 시장에서 앞서 나가고 있으나, 스타트업의 오픈소스를 토대로 한 개발 경쟁도 치열한 상황이다. 지난 2년간 AI 에이전트 관련 스타트업에 유입된 투자 자본은 20억$(약 2.78조원)를 초과하고 있다.

◇ 딜로이트 보고서(딜로이트 2025 첨단 기술·미디어·통산 산업 전망, 2025.1.23.)

- 2025년 생성형 AI를 도입한 기업 중 25%가 AI 에이전트를 시범 도입하거나 기술 검증에 나설 것으로 전망 → 2027년에 이르면 그 비율은 50%로 늘어날 것으로 예상
- 스타트업과 기존 테크 기업들이 수익 잠재력을 기대하고 AI 에이전트 개발에 힘을 쏟고 있음 → 지난 2년간 AI 에이전트 개발 스타트업에 유입된 투자 자본은 20억$가 넘음

평가·인증·프롬프트 보안·레드팀·시뮬이나 테스트데이터 등 운영·안전 시장까지 주변 생태계가 동반 확대될 것으로 예상된다. AI 에이전트는 AI 기술의 발전뿐 아니라 컴퓨팅 하드웨어, 클라우드 인프라 등 다양한 기술 생태계의 성장과도 연관되어 기술적, 경제적 성장 기회를 창출할 것으로 전망된다. 국내 AI 에이전트 관련 유망 스타트업·중소기업 등에 다양한 산업적 기회를 제공함으로써 글로벌 AI 에이전트 활용 시장을 선점할 필요가 있다.

융합 및 혁신 관련 기술의 부정적 영향

자율 에이전트의 행위와 결정 결과에 대한 법적 귀책이 모호하므로 개발·배포·운영·이용자 간 분담 설계가 필요하다. 기술 간 융합이 비표준화되어 있고, 규제 사각지대가 있으며 책임 소재가 불분명한 등 새로운 위험을 동반할 수 있다. 또 공급사에 종속되거나 프로토콜 파편화로 융합 서비스 확장을 저해할 수 있으므로 국가 표준이나 조달 요건이 필요하다.

신산업 창출이 기존 산업 생태계의 혼란이나 일자리 구조 재편을 초래할 가능성이 있다. 기업 경영에 AI 에이전트를 과도하게 의존할 경우, 편향된 의사 결정이나 윤리 이슈, 데이터 오남용 가능성도 커진다.

정책 수요와 대안

국제 표준 개발 참여와 MCP 등 사실상의 글로벌 표준(de facto)에 대응하기 위한 환경이 조성되어야 한다. MCP, A2A 등 빅테크 중심 에이전트 기술과 국내 상호 운용 생태계 구축도 필요하다. 단기적으로는 산업 현장 내 적용 유도를 위한 규제 샌드박스 적용, 법 정비 등 기업의 컴플라이언스 대응을 지원해야 하고 금융 마이데이터 등을 활용한 실시간 AI에이전트 서비스 개발을 위해 개인 정보 처리 등에 대한 현행 제도 개선이 필요하다. 개인정보보호법, 저작권법, 전자금융거래법 등, 규제 샌드박스 종료 후 지속성 등 한계를 극복하기 위한 관련법도 정비되어야 한다.

단기적으로 AI 에이전트 특화 창업팀을 발굴하고 글로벌 진출을 지원해야 한다. AI 에이전트 산업 활용 아이디어 공모전이나 CVC 등 투자 연계, 글로벌혁신센터(KIC), 외국 IT지원센터 등을 활용한 글로벌 시장 진출에 대한 지원도 필요하다.

기존 생성형 AI 시장이 AI 에이전트 시대로 급변함에 따라 기술 개발과 상용화 시장 선점을 위한 연구 개발과 실증 사업이 확대되어야 한다. 특화 파운데이션 모델, 설명 가능한 인공 지능이나 eX플레이너블(eXplainable) AI 같은 XAI, 신뢰성과 안정성 확보 기술 등을 개발하기 위한 R&D를 중기 계획으로 추진해야 한다.

다양한 산업 분야에 특화된 맞춤형 에이전트 서비스, AI 에이전트 활용 B2B/B2C 서비스에 대한 개발을 단기에 실증 지원해야 한다. AI 에이전트 개념은 오래되었으나, 산업 적용을 위한 실증은 PoC 단계로 글로벌 시장에서 경쟁이 치열해지고 있기 때문이다. 공공 분야 민원 행정 처리 등 공공 AX를 위한 AI 에이전트 서비스 지원 확대와 민간 확산 테스트 베드로의 활용도 단기 계획으로 이뤄져야 한다.

– 조명수(정보통신산업진흥원 팀장)

15. 융합 및 혁신 : 연결되는 에이전트, 복잡해지는 통제

AI 에이전트 기술은 단순 자동화를 넘어 자율적 의사 결정과 목표 달성이 가능한 차세대 AI로, 최근 서로 다른 기업과 산업의 에이전트들이 협업할 수 있는 기술적 기반이 마련되면서 여러 전문 에이전트 간 융합의 가능성이 열리기 시작했다. 2024년 말부터 앤트로픽 MCP, 구글 에이전트2에이전트(A2A) 등 빅테크의 에이전트 플랫폼과 프로토콜이 출시되며, 단일 기업 내 활용을 넘어 기업 간, 산업 간 협업의 기술적 토대를 구축하고 있다.

2024~2025년 주요 빅테크 기업들이 AI 에이전트 플랫폼과 상호운용성 프로토콜을 잇따라 출시하며, 에이전트 간 협업을 위한 기술적 인프라가 본격적으로 구축되기 시작했다. 구글의 A2A 프로토콜은 서로 다른 프레임워크로 구축된 에이전트들이 공통 언어로 소통할 수 있게 하며, 아틀라시안, 세일즈포스, SAP 등 수십 개 글로벌 기업이 파트너로 참여하고 있다. 앤트로픽의 모델 콘텍트 프로토콜(Model Context Protocol, MCP)은 AI 에이전트와 외부 시스템인 데이터베이스, 도구, API를 표준화된 방식으로 연결하며, 오픈AI와 구글 딥마인드도 이를 채택했다.

에이전트 협업 인프라를 기반으로 서로 다른 전문성을 가진 에이전트들이 복잡한 업무를 분담하는 멀티에이전트 시스템이 의료 분야를 중심으로 연구되고 시범 적용되기 시작했다. 2025년 5월 마이크로소프트가 헬스케어 에이전트 오케스트레이터를 발표했지만 이는 '연구 및 개발 목적'으로만 설계되었고 실제 임상 환경 배포용이 아니라고 명시했다. 흉부 X-ray 보고서를 생성하는 CX레포트젠, 아홉 가지 이미징 모달리티를 분석하는 메드이미지팔스, 유사 사례를 검색하는 메드이미지인사이트 등 마이크로소프트의 의료 AI 모델들을 전문 에이전트로 통합하여 협업 시스템을 구현하고 있다. 시멘틱 커넬(Semantic Kernel)의 그룹 채팅 인프라와 MCP를 활용하여 에이전트들이 "안전하고 구조화된 그룹 채팅"처럼 상호작용하는 아키텍처를 구축하고 있다.

융합 및 혁신 관련 기술의 긍정적 영향

기업 간, 산업 간 에이전트 상호운용성을 확대할 수 있다. 구글 A2A, 앤트로픽 MCP와 같은 표준화 프로토콜이 등장함으로써 특정 기업의 폐쇄적 생태계를 넘어 다양한 기업·산업 간 협업 가능성이 열렸다.

전문성 기반 에이전트 분업 체계의 실험이 가능해졌는데, 서로 다른 전문성을 가진 에이전트들이 복잡한 업무를 분담하는 멀티에이전트 시스템이 의료, 금융, 제조 등 다양한 분야에서 초기 연구가 진행되고 시범 적용 단계에 진입했다.

생태계 기반 서비스도 혁신되고 있는데 업계 표준이 마련되면, 중소 소프트웨어 기업도 글로벌 기업이 만든 에이전트와 상호 작용하며 새로운 서비스를 출시할 수 있다.

융합 및 혁신 관련 기술의 부정적 영향

에이전트 협업 인프라의 보안 취약점이 증대될 우려가 있다. 분산형 시스템의 공격이 표면으로 확장될 것이고 다수 에이전트가 상호 작용하는 복잡한 네트워크 구조 때문에 단일 취약점이 전체 시스템에 연쇄적 피해를 가져올 가능성이 커진다. 그래서 기존 보안 모델로는 대응에 한계가 있

다. 표준화 프로토콜에 의존할 위험도 있다. 특정 프로토콜에 대한 업계 의존도가 높아지면 해당 프로토콜의 장애나 보안 결함이 연결된 모든 시스템에 동시 영향을 미치는, 시스템적 위험이 커질 수 있다.

멀티에이전트 협업의 통제와 책임에도 공백이 생길 수 있다. 의사 결정 과정이 심각하게 불투명해질 것이고 여러 에이전트의 복잡한 상호 작용으로 최종 결과 도출 과정을 추적하기 어려워져 알고리즘 편향이나 오류가 발생했을 때 원인 규명이나 책임 소재를 파악하는 것이 곤란해진다. 에이전트 간 자동화된 협업 과정에서 인간의 감독과 개입이 필요한 시점을 명확히 정의하기 어려워 예기치 못한 상황이 일어났을 때 대응력이 떨어진다.

정책 수요와 대안

멀티에이전트 시스템의 분산형 구조 때문에 생기는 보안 취약점이 커지고 책임 소재가 불분명해질 것이다. 또 여러 전문 에이전트 간 복잡한 상호 작용으로 의사 결정 과정이 불투명해진다는 문제도 늘어나게 된다. 자동화된 협업 과정에서 인간이 개입할 시점이 모호해 예기치 못한 상황이 생겼을 때 대응력이 떨어질 것이다.

이런 점들을 해결하기 위해 단기적으로는 멀티에이전트 보안 프레임워크를 구축할 필요가 있다. 분산형 시스템 보안 인증, 에이전트 간 신뢰 체인 구축, 연쇄 장애 방지를 위한 격리 메커니즘 등을 의무화하는 것이다. 중장기 계획으로는 융합 시나리오 실증과 투명성 확보 체계를 구축하는 것이 필요하다. 에이전트 규제 샌드박스를 도입하여 의료법, 개인정보보호법 등을 한시적으로 완화하고 금융 AI 에이전트의 거래 한도 및 인증, 에이전트 간에 개인 정보를 공유할 때 동의 절차를 거치도록 한다. 설명 가능한 멀티에이전트 시스템을 제도화하여 에이전트 간 의사 결정 과정을 추적하고 기록 시스템을 구축한다. 또 인간이 개입할 지점을 명시화하여 투명성과 책임성을 확보한다.

– 허상우(네이버 연구원)

16. 주권 및 안보
: AI 주권이 국가 경쟁력을 결정하는 시대

AI가 국가 안보와 경제 번영을 좌우하는 전략 자산으로 떠오름에 따라 기술 경쟁력뿐만 아니라 국가 안보와 미래 번영을 위해서도 AI 주권 기반을 구축하는 것은 필수가 되었다. 한국은 AI 주권과 안보를 지키기 위해 대규모 투자와 국가 전략을 수립하여 AI 역량을 강화하고 있다. 미·중 경쟁 구도 속에서 균형을 유지하는 AI를 육성하는 정책도 추진하고 있다.

미국과 중국의 AI 패권 경쟁이 가속화되고, EU는 AI Act 제정 등 규범 정립을 주도하면서 각국이 저마다 AI 주도권 확보에 총력을 기울이고 있다. 「미국 AI 실행 계획」은 미국의 기술 주도권을 유지하고 글로벌 경쟁 구도에서 우위를 확보하기 위한 전략적 대응으로 마련된 국가 AI 전략이다. 중국의 「글로벌 AI 거버넌스 실행 계획」은 중국이 국제 규범 설정 역량을 강화하고, AI 외교를 통해 글로벌 영향력을 확대하기 위해 마련된 글로벌 AI 전략이다. 국제 협력으로 안전한 AI 개발 가이드라인을 마련하려는 움직임이 있지만 주요 강대국 간 입장 차이로 통합된 글로벌 거버넌스 수립에는 어려움이 있다.

AI 기술이 무기 자동화, 드론 정찰, 사이버전 등에 활용되며 현대전 양상을 변화시키고 있다. 따라서 각국은 안보 전략에 AI 통합을 서두르고 있다. AI 군비 경쟁은 분쟁의 예측 불가능성을 높이고 규제 공백을 드러낸다. 자율 무기의 오작동이나 오남용, AI 판단 오류로 인한 우발적 충돌 가능성 등 새로운 안보 위협과 윤리적 딜레마가 나타나고 있다.

AI 핵심 기술과 데이터 인프라가 일부 국가와 거대 ICT 기업에 집중되면서 다른 국가들은 외국 기술에 의존하게 되고 이에 따라 기술 주권이 약해질 우려가 있다. 자체 AI 모델이나 인프라가 없을 때 외국 기업의 서비스 중단이나 통제에 대응할 수 없어 자국 언어·데이터에 대한 통제권을 잃고 'AI 식민지'로 전락할 위험이 있다.

주권 및 안보 관련 기술의 긍정적 영향

글로벌 차원에서 AI 위험을 다루기 위해 진행되는 다양한 협력에 참여하여 국제 안보 강화에 기여하고 글로벌 AI 정책을 선도할 필요가 있다. 한국이 AI 윤리 기준과 기술 표준을 선도할 경우 국제 규범 형성의 주도권을 확보함으로써 소프트파워와 외교적 영향력의 증대를 기대할 수 있다.

AI 활용으로 군 의사 결정과 지휘 통제가 정밀해지고 신속해져 국방 효율성이 높아진다. AI는 사이버 안보와 방첩 분야에서도 활용되어 위협을 탐지하고 대응 능력을 강화하는 데 기여할 것이다. 미국 국방부는 AI 기반 분석으로 해킹 시도를 조기에 식별하고 방대한 보안 조사를 자동화하여 안보 태세를 높이고 있다.

AI 기술 개발을 통해 새로운 산업을 창출하고 경제 성장을 촉진할 수 있다. 스타트업, 기업 참여로 일자리와 수출이 늘어남에 따라 경제 안보도 강화하게 된다. 국가 주도로 자체 초거대 AI 모델과 AI 인프라를 구축할 경우 데이터 주권을 확보하고 종속 상황을 줄일 수 있어 기술 자립이 가능해진다. 한국은 소버린 AI를 위해 AI 연구 개발, 자체 모델 개발, 국가AI 컴퓨팅센터 구축 등을 추진하고 있다.

주권 및 안보 관련 기술의 부정적 영향

미·중 기술 냉전 속에 AI 기술의 블록화 현상이 심화되어 국제 협력이 저해될 우려가 있다. 미국은 AI 행동 계획을 통해 AI 경쟁 구도에서 미국의 글로벌 기술 리더십을 강화하기 위한 '풀스택(full-stack) AI 수출 패키지' 전략을 발표했다. 각 블록이 자국 우선으로 움직이며 핵심 기술과 데이터의 디커플링(탈동조화)을 초래하면 글로벌 공급망과 기술 표준의 단절이 우려된다. 통일된 AI 거버넌스가 없을 때 국제 표준 미비로 AI 윤리·안전 확보에 공백이 발생할 수 있다. AI 규제에 대해 국가별로 서로 다른 접근으로 일관된 글로벌 규범 형성이 어려워 AI 위험에 대한 국제 공조 대응이 지연되고 있다.

예측 불가능한 AI 알고리즘이 통제되지 않을 경우 의도치 않은 분쟁으로 확산될 수 있으며 이는 군사적 위기 관리의 어려움을 더하게 된다. AI를 악용한 사이버 공격과 정보전 위협이 증가하여 국가 안보와 사회 안정에 새로운 위험 요소로 등장하고 있다.

특정 국가 기술에 대한 과도한 의존은 위기가 발생했을 때 심각한 취약점을 드러내게 된다. AI 칩이나 소프트웨어에 대한 수출 통제 등이 현실화되면 해당 기술에 의존하던 국가의 AI 인프라가 마비되어 안보·경제에 큰 타격을 입을 수 있다. AI 기술 패권국이 '신(新)제국주의' 수단으로 AI를 활용해 다른 나라의 데이터와 디지털 영역을 지배할 가능성도 있다. 기술 종속 국가는 산업 생태계까지 외부 영향권에 놓여 자국 주권과 자율성이 심각하게 흔들리는 위험에 처하게 된다.

정책 수요와 대안

국경을 초월한 AI의 영향에 대응하여 디지털 주권을 지키고, 국제 AI 규범을 만드는 데 한국이 주도권을 확보하기 위해 국제 협력을 이끌어가야 한다. 글로벌 협업을 통해 AI 에이전트의 데이터 처리와 알고리즘 투명성, 상호운용성 등에 대한 공통 규범을 마련해야 한다. G7, G20, UN 등 주요 국제기구에서 AI 에이전트 운용 원칙과 상호운용성 표준을 선제적으로 제안할 필요가 있다. AI 에이전트의 오용을 방지하기 위한 국제 협약을 체결하기 위해 노력해야 한다. AI 안전성과 신뢰성을 검증하는 기술을 공동으로 연구할 필요도 있다.

글로벌 AI 경쟁에 대응해 AI 유니콘 육성과 오픈 소스 생태계 활성화를 통해 기술 종속을 줄이고 국가 AI 경쟁력을 높여야 한다. AI 에이전트 분야 예비 유니콘을 육성하기 위한 대규모 투자 펀드를 조성하고 규제 샌드박스를 확대하는 등 AI 유니콘 기업 창출 생태계도 조성해야 한다. 공공부문 우선 도입을 통해 국산 AI 시장을 창출하고 중소기업 도입 지원금을

제공하여 내수를 확대해야 하며 오픈 소스 AI 생태계 활성화와 국산 AI 기술의 외국 의존도를 단계적으로 줄여나가는 로드맵을 만들어 AI 기술이 종속되는 것을 막는 기반을 구축해야 한다.

AI 에이전트의 자율적 사이버 공격과 무기화로 인한 새로운 안보 위협과 군비 경쟁에 대응하고, 국제 안보 불안을 해소하는 데 힘써야 한다. 국방 AI 역량을 강화하고 민관군 협력을 통해 사이버 위협에 대응하며 자율무기 국제 규제를 선도해 나갈 필요가 있다. 민간 사이버 보안 기업과 국방부 간에 실시간으로 위협 정보를 공유하는 네트워크를 운영하고 AI를 기반으로 한 위협을 예측하는 시스템을 도입해야 한다. 완전 자율 무기 체계(LAWS) 개발 금지와 인간 통제하에 AI 무기 시스템만 허용하는 국제 협약 추진도 검토해야 한다. 국제인도법 준수를 의무화하고 AI 무기를 사용했을 때 지휘관 책임을 명확히 하여 전쟁 범죄를 막는 체계를 구축해야 한다.

AI 안전성과 윤리의 국제 표준을 선도하고, AI 에이전트의 책임 있는 활용을 위한 법적·제도적 체계를 만들 필요가 있다. 모든 AI 에이전트를 의무적으로 등록하는 제도를 도입하고 행위 로그 의무 보관을 통한 추적 가능성을 확보하는, AI 에이전트 행위 체계를 세워야 한다. AI 에이전트의 자율성을 수준별로 나누어 차별적으로 규제하는 체계를 도입하고 위험성이 높은 AI에 대해 인간이 의무적으로 개입하도록 해야 한다.

감시 오남용과 자율 무기 인권 침해를 방지하고, 시민의 안전과 권리를 보호해야 한다. 최종 판단은 인간이 수행하는 인간 중심 AI 보조 감시 시스템 개발이 필요하고 프라이버시 보호 기술도 적용되어야 한다. AI 감시 대상과 범위를 법으로 엄격히 한정하고, AI 감시 데이터 보관을 제한 등 AI 감시 시스템 오남용을 방지해야 한다. 자율 무기를 사용할 때 민간인을 보호하는 의무를 강화하고 AI 표적 식별 시스템의 오인식률을 최소화하는 기술을 개발해야 한다.

빅테크의 AI 기술 독점에 따라 산업이 종속되고 기술 격차가 심화되는 것을 막고, 국가 경쟁력 저하를 방지하기 위한 전략적 대응이 필요하다. 이를 위해 국가 AI 챔피언 기업을 육성하고 국제 표준화와 공정 경쟁 규제를 통해 한국의 AI 주도권과 기술 경쟁력을 강화해야 한다. AI 국부 펀드를 조성하고 외국 AI 기업 M&A 지원, 글로벌 AI 인재 유치 등을 통해 국가 AI 챔피언 기업을 육성하는 데 힘써야 한다. AI 표준 특허 확보를 지원하고 특허 풀 구축을 통한 기술 경쟁력을 강화할 필요가 있다.

반독점 규제를 강화하고 데이터 공유와 상호운용성을 표준화하며, 공공 컴퓨팅 자원 지원 등을 통해 AI 시장 독과점을 방지해야 한다. 반독점 집행과 상호운용성을 의무화하고 데이터 이동성을 확보하여 지배적 외국 기업의 통제를 견제하고 개방적이고 경쟁적인 AI 생태계를 조성할 필요가 있다. AI 시장 점유율 상한제를 도입하고 핵심 AI 기술에 대한 강제 라이선싱 제도을 도입하여 AI 시장 독과점을 방지해야 한다. 국제 수준의 데이터를 공유하고 상호운용성 표준화를 통해 공급업체 종속을 줄이고 다양성과 지역 대표성을 갖춘 AI 데이터셋을 확보할 필요가 있다. 스타트업과 중소기업을 대상으로 고성능 컴퓨팅 자원을 저렴하게 제공하는 공공 AI 클라우드와 슈퍼컴퓨팅 인프라 구축도 검토해야 한다.

– 김태원(한국지능정보사회진흥원 수석)

17. 주권 및 안보 : AI 안보 시대의 통제와 책임

AI 에이전트 기술은, 공공 및 안보 분야에 도입되면 국가 운영의 효율성과 자주성을 높일 수 있는 잠재력을 지니고 있으며, 2024년부터 제한적이지만 실질적인 적용 사례들이 나타나기 시작했다. 공공 부문에서는 복잡한 행정 절차를 자동화하고 부처 간 정보를 연계하며 긴급 상황 대응 등에서 AI 에이전트의 가능성이 주목받고 있다. 그러나 높은 신뢰성이 요구되고 책임 소재 문제로 인해 대부분 국가에서 제한적 시범 사업 수준에 그치고 있다.

국방·안보 분야는, 미국에서 스케일 AI의 군사 계획 지원 AI 에이전트(Thunderforge) 시범 도입하는 등 특정 영역에서 성과를 보이고 있다. 하지만 자율 무기 체계나 완전 자율 작전 시스템으로의 확장은 통제 가능성과 윤리적 우려로 인해 매우 신중하게 접근하고 있다. 특히 글로벌 빅테크가 주도하는 AI 에이전트 플랫폼의 표준화 추세는 기술 도입의 편의성을 제공하는 동시에, 국가 핵심 기능이 외부에 종속되고 데이터 주권이 침해된다는 새로운 도전 과제를 제시하고 있어, 각국은 기술 혁신과 주권 보호 사이의 균형점을 모색하는 상황이다.

AI 에이전트 기술이 공공 행정에 실험적으로 도입되기 시작하면서 정책 집행의 효율성과 주권적 의사 결정 능력이 동시에 시험대에 오르고 있다. 자체 AI 시스템을 구축할 때 정부의 독자적 정책 수행 능력과 국민 데이터 주권을 확보할 수 있다는 기대가 높아지고 있다. 하지만 에이전트의 자율성 증가에 따른 민주적 통제가 약해지고 외국 플랫폼을 의존할 때 주권이 침해될 수 있다는 우려가 공존한다.

AI 에이전트의 군사적 활용이 방어 역량을 혁신적으로 향상시킬 수 있으나, 동시에 예측 불가능한 자율 시스템 간 상호 작용과 새로운 형태의 안보 위협을 초래하고 있다. 24시간 무인 국경 감시나 실시간 위협 탐지 등을 통해 비대칭 위협에 대한 효과적 대응이 가능해지고 있다. 그러나 AI 군비 경쟁이 가속화되고 자율 무기를 통제할 수 없다는 점에 대한 국제 사회의 우려도 커지고 있다. 특히 적대국의 AI 시스템을 해킹하거나 기만, 조작 등 기존에 없던 사이버-물리 융합 위협이 등장하면서 새로운 안보 패러다임 정립의 필요성이 대두되고 있다.

주권 및 안보 관련 기술의 긍정적 영향

공공 부문에 AI 에이전트를 도입하고 통제하면 정책 집행력 강화를 통한 실질적 주권 행사 역량을 높일 수 있다. AI 에이전트의 자율적 정보 수집·분석·대응 능력으로 복잡한 정책 환경에서도 일관되고 신속하게 국가 의사 결정이 가능하다. 자체 AI 시스템을 구축하면 외국 플랫폼 의존 없이 국민 데이터의 주권적 활용과 보호를 동시에 달성할 수 있다.

국방·안보 분야에 자율 시스템을 활용하면 비대칭 위협 대응과 자주 국방 역량을 강화할 수 있다. 24시간 무인 감시, 신속한 초동 대응으로 소규모 병력으로도 효과적 방어 태세를 구축할 수 있다. 다양한 국가나 기관 에이전트 간 자동으로 실시간 정보 교환이 용이해져, 테러·사이버 공격 등 글로벌 위험에 대해 연합·공조 역량이 강화될 수도 있다.

주권 및 안보 관련 기술의 부정적 영향

공공 부문에 AI 에이전트를 도입하면 의사 결정 주권을 잠재적으로 상

실할 위험이 있다. 에이전트의 자율적 판단에 과도하게 의존할 경우 인간의 정책 결정 역량이 약화되고 민주적 통제 기능이 훼손될 우려도 있다. 복잡한 멀티에이전트 시스템에서 오류가 발생했을 때 원인을 파악하거나 책임 추적이 어려워 거버넌스 체계가 약화될 수 있다. 외국 AI 플랫폼을 사용할 때 정책 결정 패턴과 국가 운영 데이터가 학습되어 주권적 의사결정을 침해할 수도 있다.

국방·안보 분야에 자율 시스템을 활용하면 새로운 안보 딜레마에 빠지고 통제권을 상실할 위험이 있다. 자율 시스템 간 예측 불가능한 상호 작용으로 인간이 통제할 수 없는 확전 위험도 발생할 수 있고 AI 군비 경쟁 가속화로 인해 지역 불안정성이나 우발적 충돌 가능성이 커진다. 또한 적대국의 AI 시스템을 해킹하거나 기만, 조작하는 등 새로운 형태의 안보 위협이 출현할 가능성이 있다. 또한 자율 무기 체계의 생명에 대한 결정권 문제와 국제인도법 위반도 우려되며 오작동이나 오판으로 인해 민간인이 피해를 입었을 때 책임 소재가 불명확하다는 문제도 생긴다.

정책 수요와 대안

AI 에이전트를 활용한 자율 감시, 사이버 위협 탐지 등이 방어 역량을 향상시키고 있다. 하지만 자율 시스템 간 예측 불가능한 상호 작용으로 인한 확전 위험과 적대국의 AI 시스템 해킹 등 새로운 형태의 안보 위협이 증가하고 있다.

이 문제를 해결하기 위해 군사용 AI 에이전트 안전 통제 체계를 구축하고 자율 무기 체계와 AI 에이전트의 생명 관련 결정에 대한 엄격한 통제 기준을 마련해야 한다. 자율 무기 체계에 대해서는 반드시 인간의 감시, 통제 원칙을 적용하는 것을 의무화한다. AI 에이전트 간 상호 작용으로 인한 우발적 확전을 방지하기 위한 '킬 스위치(kill switch)'나 긴급 정지 시스템을 의무적으로 탑재하는 것도 검토해야 한다.

– 허상우(네이버 연구원)

AI 생성 이미지

18. AI의 지속 가능성 : 탄소중립 시대의 AI 딜레마

AI 에이전트는 기존 LLM, 생성형 AI를 넘어서는 AI로, 막대한 연산 자원과 전력을 필요로 하며 탄소 배출과 환경 오염을 유발한다. AI 에이전트가 산업·공공·교육 현장 전반에 확산될수록 환경 부담과 기술·사회 구조의 지속 가능성 확보가 핵심 과제가 될 것이다. 국제적으로 그린 AI(Green AI)와 지속가능한 AI(Sustainable AI) 논의는 OECD, EU, UN 등을 중심으로 점차 활발해지는 추세이다. 미국과 북유럽은 데이터센터에 재생 에너지와 해양, 풍력 등 자연 냉각 시스템을 도입 중이며 구글, MS, 메타, 오픈AI 등 주요 AI 빅테크들은 모델 경량화와 에너지 효율화 기술 개발을 추진하고 있다.

LLM 학습·운영 시 많은 전력이 소모되며, AI 에이전트는 지속적 실행을 해야 하므로 상시 전력 소모가 크다. 대개 AI의 학습과 추론 시에는 막대한 데이터와 계산 자원이 필요하며, 이를 위한 인프라인 데이터센터는 상당량의 전력을 소모한다. 또한 데이터센터의 냉각과 전력 공급(생산) 과정에서 다량의 탄소가 배출되므로 전력 소모를 줄일 수 있는 연산 효율성을 지닌 경량 AI 모델 개발이 필요하다.

지속 가능한 AI 에이전트를 담보하기 위해서는 데이터 수집·활용의 합법성·윤리성이 확보되어야 한다. 편향된 데이터 사용은 기술 신뢰성을 저하시키고 사회 지속 가능성에 악영향을 미친다. 또한 대규모 AI 인프라를 보유한 소수 기업이 기술을 독점하면 AI 혜택에 대한 편중과 독과점이 발생할 수 있다.

AI의 지속 가능성 관련 기술의 긍정적 영향

AI 에이전트를 ESG 달성과 탄소 배출을 줄이는 시스템에 활용하여 환경 오염을 방지하는 효과를 높일 수 있다. AI 에이전트 기술 개발 과정에서 NPU, TPU과 같은 저전력 AI 칩, 엣지컴퓨팅, 지능형 자원 관리 등의 기술이 활성화되고 있다. AI 에이전트를 활용하여 에너지 생산이나 제조 과정에서 탄소 배출량에 대한 실시간 모니터링과 자동화된 보고 체계 등을 갖추는 것이 가능해진다.

제조·물류·건설 등 산업에서 AI 에이전트가 자원 사용 최적화와 폐기물 감소에 기여한다. 기술 고도화로 인한 장비 교체 주기 단축되어 전자 폐기물이 증가하는데 이를 다양한 산업에서 재활용할 수 있다.

AI의 지속 가능성 관련 기술의 부정적 영향

AI는 학습할 때 방대한 컴퓨팅 자원을 필요로 하며 필연적으로 많은 양의 전력을 소모한다. 24/7 모니터링, 실시간 대응 등 AI 에이전트를 상시 운영하면 전력 사용을 고정적으로 높일 수 있다.

대규모 AI 인프라 구축이 가능한 기업과 그렇지 못한 기업 간 기술 격차가 심화될 수도 있다. 고성능 AI 컴퓨팅 자원은 초기 투자비가 수천억~수조 원 규모에 달해, 자금력이 부족한 중소기업들의 인프라 구축은 사실상 불가능할 수 없다. 특정 기업이 AI 컴퓨팅 인프라나 데이터 자원을 독점하면 AI 접근성과 지속 가능성을 떨어트릴 수도 있다.

정책 수요와 대안

AI 에이전트의 높은 연산 비용과 에너지 소모로 인한 환경적 지속 가능성 우려를 해소하기 위해 엣지 컴퓨팅 등을 활용한 저전력·경량 AI 모델 개발을 장려하고, AI 모델의 효율성을 측정하는 라벨링 제도 도입을 고려할 만하다. 가칭 '그린 AI 인증제'라 하는 저전력·친환경 저탄소 모델 인증 제도를 추진하고, 인증받은 모델에 한하여 공공 조달을 가능하게 하거나 국내 보급 확산을 지원하는 것도 방법이다. AI 데이터센터의 전력 소비량을 줄일 수 있는 관리 최적화 시스템과 학습할 때 전력 소모를 줄이는 알고리즘 R&D 지원 사업이 추진되어야 한다.

태양광, 풍력, 조력 등 신재생 에너지 개발을 고도화하고 확산 및 폐기물 재생(RE 100) 등을 통해 넷제로(Net-zero)를 달성해야 한다. AI 데이터센터에 공급하는 전력은 친환경 방식으로 생산된 전력만 활용할 수 있도록 인센티브 법안을 추진할 수도 있다. 에너지 부문이나 산업별 AI 주요 리더들로 구성된 'AI 에너지 위원회' 등 컨트롤 타워를 설립하여 저탄소 에너지원 활용 방안을 모색해야 한다.

– 안성원(소프트웨어정책연구소 실장)

19. AI의 지속 가능성 : 지속 가능한 AI 인프라의 가능성

AI 에이전트는 사용자의 지시 없이도 목표 지향적으로 환경과 지속 상호 작용하며 계획 수립·도구 사용·피드백 학습을 반복 수행하는 체계이다. 다만 상시 동작이나 지속 통신 구조의 AI 에이전트 특성으로 인해 연산·저장·네트워크 사용이 누적되어 전력·냉각·물·자원 수요가 함께 커지는 속성을 지니고 있다. 사회·윤리·법 영역에서는 장기 상호 작용으로 축적되는 개인 정보·행동 로그에 대한 보호·감사 가능성과 책임 귀속의 명료화가 동시에 요구된다. 에너지 측면에서는 신재생 연계 데이터센터, 탄소 인식 스케줄링, 액체 냉각·폐열 회수 등 고효율 인프라 전환이 가속화될 것으로 내다볼 수 있다. 국제 사회에서는 AI의 환경적 지속 가능성과 포용적 접근, 그리고 이를 위한 구체적 실행 방안을 제시하며 AI 발전 방향에 중요한 전환점을 제시하고 있다.

산업계는 대규모 모델을 경량화·선택적 활성화 모델(MoE)·온디바이스 분산으로 전환하며 비용과 전력의 최적화를 시도 중이다. 데이터센터는 액체 냉각·폐열 회수나 컴퓨팅 부하를 전력의 탄소 강도를 고려해 스케줄링하는 방식인 탄소 인식 스케줄링 등 고효율 설계를 채택하는 추세이다. 다만 전력망 수요가 급증하고 희소 금속 공급망, 사용 후 장비의 회수나 재제조 등 시스템 차원의 병목이 두드러진다.

유럽 중심으로 규제·표준 측면에서는 환경 정보 공개, 그린 구매, 데이터 최소화, 폐기 프로토콜 요구가 강화되는 흐름이다. 국제 협력 의제는 카본 어카운팅 방법론 통일, 그린데이터센터 기준, 폐기 인증으로 수렴되는 중이다.

AI 지속 가능성 관련 기술의 긍정적 영향

전력·수열·가스 등 에너지 시스템에서 AI 에이전트가 수요 예측, 부하 이동, 피크 제어를 자율 수행함으로써 재생 에너지 수용성을 확대하고 계통 안정화에 기여할 가능성이 있다. 환경·자연 자원 관리에서 위성·IoT·현장 센서와 상호 작용하는 에이전트가 산불·홍수·적조 등 위험 조기 경보, 수자원 배분 최적화, 농업 투입재 정밀화를 달성할 수 있으며, 이는 자원 투입량 절감이나 폐기물 최소화, 복구 비용 절감으로 연결된다.

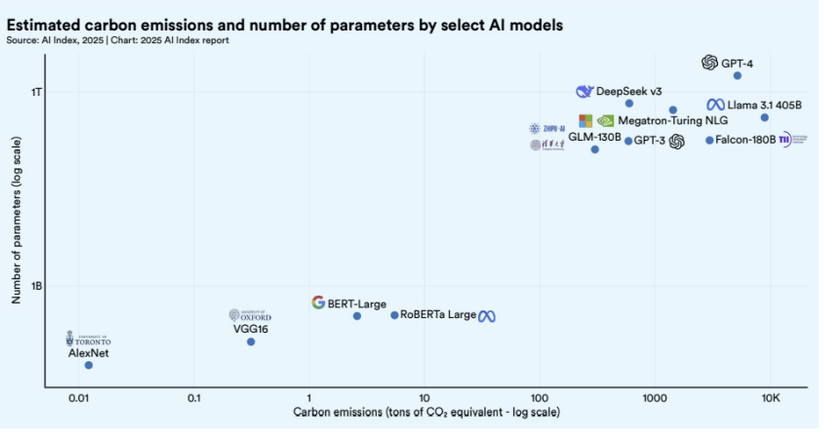
제조·물류 분야에서 공정이나 라인·설비 단위의 다중 에이전트 협업이 품질 편차를 축소하고 비계획 정지를 감소하며 이는 에너지 단가 하락으로 이어진다. 데이터·모델·행동 로그가 자산화되면서 에이전트-애즈-어-서비스(AaaS), 그린 옵스 BPO(Green Operation Business Process Outsourcing, 기업의 특정 업무 프로세스를 외부 전문업체에 맡김) 등 신규 시장 형성이 기대된다. 금융이나 헬스·공공 서비스에서 지속 상호작용형 에이전트가 맞춤 의사 결정 지원을 제공함으로써 서비스 접근성 확대와 운영비 절감과 함께 실시간 정책 집행 효율이 향상된다.

AI 지속 가능성 관련 기술의 부정적 영향

AI 에이전트 상시 운영으로 기저 전력 수요가 상승될 수 있으며, AI 모델·전력 인프라 수준 차이에 따라 탄소 집약도 증가할 수 있다. AI 모델 규모가 확대되면서 탄소 배출 증가가 늘어나고 있다. 챗GPT 3 모델의 경우 학습 단계에서 CO2가 503t 배출되는 것으로 보이며 최근 모델인 챗GPT 5 모델은 수배 이상 더 높을 것으로 추정된다. AI 에이전트 모델의

경우, 지속적으로 상호 작용하고 여러 작업을 수행하며 일반적 모델보다 리소스를 더 많이 소모할 가능성이 있다.

〈AI 모델별 훈련 탄소 배출량〉



* 출처 : Artificial Intelligence Index Report 2025(Stanford Univ.)

장기 상호 작용 과정에서 센서 오류나 편향 데이터가 누적되면 오류의 자동화가 발생할 수 있으며 의료·재난·금융 등의 크리티컬 도메인에서 거짓 양성·음성의 시스템적 확대가 일어날 수 있다. 멀티 에이전트 시스템은 개별 에이전트 간 상호 작용을 통해 편향이 강화되거나 증폭될 수 있는 잠재적 구조 취약성을 지니고 있다.

개인화 서비스 과정에서 개인 정보 침해·유출 위험이 증가하므로 데이터 최소화나 차등 프라이버시·연합 학습 등 보호 기제가 필요하다. 종료·재설정 부재 시 유령 에이전트가 남아 있을 수 있으며, 행동 로그나 지식 베이스에 대한 돌이킬 수 없는 삭제 실패가 프라이버시와 영업 비밀 침해로 이어질 위험이 크다. 또 책임 주체가 불분명할 때 법적 분쟁 비용이 커질 수도 있다.

정책 수요와 대안

지속 가능한 AI를 위한 국제 협력 활동을 강화하고 어젠다를 주도해야 한다. 중기 계획으로, 저탄소 AI 기술이나 국내 실증 사례 공유 등을 위한 글로벌 협력 프로젝트를 추진하고 국제 공동 연구 플랫폼 참여를 확대해야 한다.

기후테크 스타트업을 육성하고 탄소 중립 실증을 확대하는 등 산업 기반을 조성해야 한다. 인공지능을 활용한 기후 AI 스타트업을 발굴·육성하는 일에 정부 지원을 강화하는 것도 필요하다. 현시점에 탄소 절감을 위한 전략망을 관리하고, 스마트 에너지 등 탄소 감축 분야 스타트업을 집중 육성하는 것도 중요하다. 이를 위해 모태펀드 내 기후테크 등을 위한 전용 펀드 구성이나 기후 스타트업 지원 트랙을 생성하는 것을 검토할 만하다.

또 현시점에 AI 반도체나 서버 등 보급 중심에서 순환 경제 달성을 위한 자원 재사용과 재제조 기술, 폐기 후 재활용 기술 등 연구 개발을 추진해야 한다. 이는 친환경 데이터센터 구축을 위한 재생 에너지 연계형 인프라 지원이나 액체 냉각, 폐열 회수 등 고효율 기술에의 적용 확대를 포함한다.

현시점에 AI 에이전트 등을 활용한 탄소 중립 제조 공정을 실증 지원하여 탄소 감축, 에너지 사용 효율 개선과 공정 자동화 수준을 높일 수 있다. 비R&D 분야인 이 사업명은 'AI 에이전트 활용 탄소 중립 제조 공정 실증 지원'이라 부를 수 있다.

중기 계획으로는 다양한 에너지원 확보와 관련 산업 촉진을 통해 지속 가능한 AI 생태계를 조성할 수 있다. 재생 에너지, SMR 등 다양한 에너지원과 분산 전원을 확보하여 AI 확산으로 인한 수요에 대응할 필요가 있다.

탄소 절감을 위한 전략망을 관리하고, 스마트 에너지 등 탄소 감축 분야 스타트업을 현시점에 집중 육성하는 것도 중요하다.

AI 모델의 에너지 저감과 AI 에이전트 활용으로 인한 다른 분야 산업 효율 향상을 중기 계획으로 확대할 수 있다. MoE(Model of Experts) 모델 개발, NPU 기반 추론 시스템 구축, 양자화 기술, 스파스(Sparcity) 모델링 기술 개발 등으로 AI 에너지 절감을 확대할 수 있다. 소규모 파운데이션 모델, 신경망 구조 혁신, 하이브리드 AI, 데이터 선택적 학습 등 에너지 절감을 위한 R&D와 실증 지원이 필요하다.

– 조명수(정보통신산업진흥원 팀장)

20. 책임 : 판단하는 AI 시대의 책임 문제

AI 생성 이미지

AI 에이전트는 자율적으로 목표를 세분화하고 외부 자원을 활용하여 행동을 수행하는 특성을 지니고 있다. 그 결과 발생하는 행위와 영향에 대한 책임 귀속 문제가 불가피하게 제기될 수밖에 없다. 현행 법 제도는 주로 개발자·운영자·사용자 등 인간 중심의 법적 책임 관념에 기반해 작동하므로, AI 에이전트의 행위 결과에 대한 책임 주체는 불명확한 실정이다. 특히 의료·군사·금융 등 고위험 분야에서 사고가 발생했을 때 책임 소재가 불명확하고 이는 사회적 신뢰와 안전성에 직접적 위협이 되고 있다.

최근 외국과 국내에서 일어난 사례에서 책임 불명확성이 사회적 논란으로 이어지고 있다. 2025년 16세 청소년의 자살 사건과 관련하여 부모

가 오픈AI 상대로 '자살 조장·방치' 소송을 제기했다. 2022년에 일어나고 2024년 판결된 에어 캐나다 챗봇 사건은 AI 챗봇의 잘못된 환급 정보 제공에 대해 법원이 회사 책임을 인정하여 보상 명령을 내린 사건이다. 2025년에 일어난 오퍼레이터 사건은 오픈AI의 '오퍼레이터' 에이전트가 승인 없이 장보기 주문을 실행한 사건이다. 사전에 구매 확정 절차를 거친다는 안전 장치가 있었음에도 이를 우회한 것으로 알려져 있다. 2019년에 일어나고 2025년 판결된 테슬라 자율 주행 사고는 자율 주행 모드에서 발생한 사망 사고에 대해 미국 플로리다주 배심이 운전자 과실과 제조사 책임을 병행 인정한 사건이다.

AI 에이전트의 사용 범위를 규정하고, 사고가 발생했을 때 민사·형사 책임 주체를 명확히 하기 위한 다양한 입법적 검토가 진행 중이다. EU는 2024년 최종 통과된 'AI Act'에 고위험 AI의 책임과 규제 방안을 포함시켰다. EU는 2024년 AI로 인해 발생한 손해에 대한 제공자와 운영자의 책임을 강화하고 피해자를 보호하려는 AI 책임 지침(AI Liability Directive, AILD)를 검토했으나, 2025년 EU 집행위에서 최종적으로 추진을 중단하였다. 미국은 일정한 투명성 요건을 충족시킨 AI 개발자를 AI로 발생한 사고의 책임으로부터 면책시키는 RISE(Responsible Innovation and Safe Expertise) Act를 2025년 6월 상원안으로 발의했다. 한국은 AI 운영 투명성·위험관리·윤리 원칙을 규정한 AI 기본법 제정하여 2025년 공표, 2026년 시행 예정이다. 다만, 책임 귀속 메커니즘은 미비한 상황이다.

AI 에이전트 활용에 대한 책임 정비가 이루어질 경우, 기술에 대한 사회적 신뢰와 수용성을 높일 수 있다. 하지만 사용자 보호, 피해 구제, 사고 대응 체계가 뒷받침되지 않을 경우, 기술에 대한 사회적 신뢰와 수용성이 훼손될 우려가 있다. 예컨대, 자율 주행 시 일어난 사고를 둘러싼 책임 귀속 문제의 불투명성은 자율 주행 서비스 확대에 어려움을 불러오고 있다.

책임 관련 기술의 긍정적 영향

현재 인간 행위자 중심의 단층적 책임 논의에서 벗어나, 비인간 행위자까지 포괄하는 다층적 법적 책임 체계와 사고 예방·대응 방안이 마련될 수 있다. 사회보험, 민간보험, 면책 특권, 민사·형사적 처벌 등을 활용한 다층적 책임 체계가 마련되고 전담 기구가 설립되면 사회적 신뢰를 높일 수 있다.

책임 귀속 논의를 통한 투명성과 설명 가능성을 강화할 수 있으며 책임 분담 요구는 투명한 개발·운영 기준 마련으로 이어질 수 있다. 입증 책임을 피해자로부터 개발자·운영자·사용자로 돌려 AI 운영의 투명성 강화하고 설계 단계에서 안전성과 설명 가능성을 확보하는 기술 발전을 이끌어낼 수 있다.

책임 구조 확립은 산업적 예측 가능성을 높여 안정적 투자·활용 환경을 조성 새로운 산업 분야를 창출할 수 있다. 이는 AI 에이전트 사고 관련 보험·인증·감독 등 연계 산업 활성화로도 이어질 가능성이 있다.

책임 관련 기술의 부정적 영향

자율적 판단과 행위로 인한 결과가 인간 주체의 의도와 무관할 경우, 책임 회피나 전가를 할 가능성이 있다. 이런 경우 개발자-운영자-사용자 간 분쟁이 심화될 수도 있다.

의료·군사·금융 등 분야에서는 오류나 오작동이 일어날 때 막대한 피해가 발생할 수 있다. 이와 같은 고위험 분야에서의 활용은 사회적 비용을 늘어나게 한다.

AI가 사용자 의지에 반하는 결정을 내려 사고가 발생할 경우, 윤리적·법적 충돌이 발생할 수 있다. 개별 혹은 집단 AI 에이전트의 효율성 추구가 사회 윤리·환경 윤리나 공공선을 훼손할 경우, 책임 소재 문제가 발생한다. 책임 주체가 불명확할 때 피해자 구제가 늦어지고 사회적 혼란이 발생할 수 있다. AI 모니터링 에이전트(AI monitoring agent) 등 감시형 기술의 도입은 사회 통제와 관련된 논란을 야기할 가능성도 있다.

정책 수요와 대안

종합적 정책 방향 : AI 에이전트의 사고와 피해에 대한 대응은 단일 제도로는 한계가 있으며, 시민보험–민간보험–법적 책임으로 이어지는 다층적 보장 체계가 필요하다. 가벼운 피해는 시민안전보험 같은 보편적 사회보험으로 신속하게 처리하여 사회적 불안을 최소화할 필요가 있다. 중범위 피해의 경우에는 민간보험을 활용하여 개발사·운영사·사용자가 책임을 분담하도록 하고, 보험시장의 활성화를 통해 새로운 산업 기반 마련을 유도해야 한다. 대규모 피해나 인명 피해 같이 중대한 사건에 대해서는 법적 책임을 강력하게 적용함으로써 최종 안전망을 확보해야 한다. 동시에 이러한 책임의 단계화·다층화 과정이 전체 AI 산업 발전을 위축시키지 않도록 입법 단계에서 조건부 면책 요건과 절차를 명확히 하여, 선의의 준수자에게 예측 가능성을 부여할 필요가 있다.

인간 중심의 현행 법제도 하에서 AI 에이전트의 자율적 판단과 행위로 인해 발생하는 결과에 대한 책임 주체가 불명확한 점을 해소할 필요가 있다. 이는 면책 인센티브 제공을 통한 산업 발전 촉진과 병행할 수 있다.

이에 대한 대안으로 AI 에이전트 활동에 대한 책임 분담 체계 수립하는 것을 들 수 있다. 개발 단계에서 투명성과 설명 가능성 확보를 위한 설계를 의무화하고 사고가 발생했을 때, AI에 관한 전문 지식을 갖추지 못한 피해자가 아닌 개발자·운영자·사용자가 책임 설명을 입증하도록 전환해야 한다. 일정 수준의 안전성과 투명성을 확보한 경우, 개발자나 운영자에게 조건부 면책을 제공하는 등 인센티브 제도를 도입하는 것도 검토할 만하다. 분쟁 중재를 위한 〈AI 피해 분쟁위원회〉를 설치하고 운영할 수 있다. AI 에이전트의 법적 지위 혹은 법인격 부여 여부에 대한 연구 활동 지원도 필요하다. 이는 장기적 관점에서 법인격 확대 여부를 결정해야 한다.

의료, 군사, 금융 등 분야에서 AI 에이전트의 오류 또는 오작동이 발생했을 때 피해가 확산되는 것을 방지해야 한다. 이를 위해 고위험 분야 AI 서비스에 대한 사전 규율과 단계적 보장제도를 연계할 필요가 있다. 분야별 위험도 평가 기준을 수립하고 인증제·허가제 등의 도입을 검토할 만하다. 고위험 서비스 제공자는 민간보험 가입을 의무화하여 사회보험 보장범위를 넘어서는 중범위 사고에 대비하게 하고 대규모 사고나 인명 피해가 발생했을 때는 강력한 형사·민사의 법적 책임을 병행하여 최종 안전망을 마련해야 한다.

AI 에이전트의 사고가 발생했을 때 체계적인 피해자 보호와 신속한 보상이 필요하다. 이를 위해서는 단계적·다층적 보장 체계를 확립해야 한다. 가벼운 경미한 피해는 현재 운영 중인 지자체 시민안전보험에 AI 에이전트 관련 사고를 포함시켜 신속히 구제하고 중범위 피해는 AI 개발사·운영사의 보험 가입을 의무화, 사용자의 민간 보험 가입 권유를 통해 보상 책임 분배가 이루어지도록 유도해야 한다. 대규모·중대 피해에 대해서는 법적 책임을 강화하여 개발사·운영사에 대한 형사·민사 제재 부과를 검토할 만하다.

사고 대응 및 피해자 보호 제도 운영이 필요하다. AI 관련 사고의 보험

통계(actuarial statistics) 자료를 신속하게 구축하기 위해 정부 차원의 AI 에이전트 사고 보고 체계를 구축해야 한다. 'AI 피해분쟁위원회' 같은 전담 기구가 피해를 접수하여 배상 기준을 판단하고 보험·법적 절차를 연계하는 등 통합 관리가 필요하다. AI 모니터링 에이전트 운용을 통한 사고 예방과 감시 권한 남용 방지를 위한 규제를 병행해야 한다.

다층적 보장 구조를 제도화하기 위해 단기·중기·장기 단계별 과제를 구분할 필요가 있다. 단계별 과제의 정책 대안은 아래와 같다.

*** 단기 과제: 현행 법제 보완**

- 민법·형법 해석 확장: AI 에이전트 사고를 불법 행위·업무상과실로 포섭
- 입증 책임 전환: 피해자가 아닌 사업자가 자료 제출·과실 추정에 대응하도록 규정
- 지자체 시민안전보험에 AI 사고 보장 항목 포함

*** 중기 과제: 특별법 정비**

- AI 기본법 보완: 안전성·투명성·책임성 원칙 명문화 및 '피해자 보호장' 신설
- 금융분쟁조정위 모델을 참고한 〈AI 피해 분쟁조정위원회〉 설립 운영
- 고위험 AI 서비스에 대한 인증·감독제 도입 및 민간보험 의무 가입 제도화

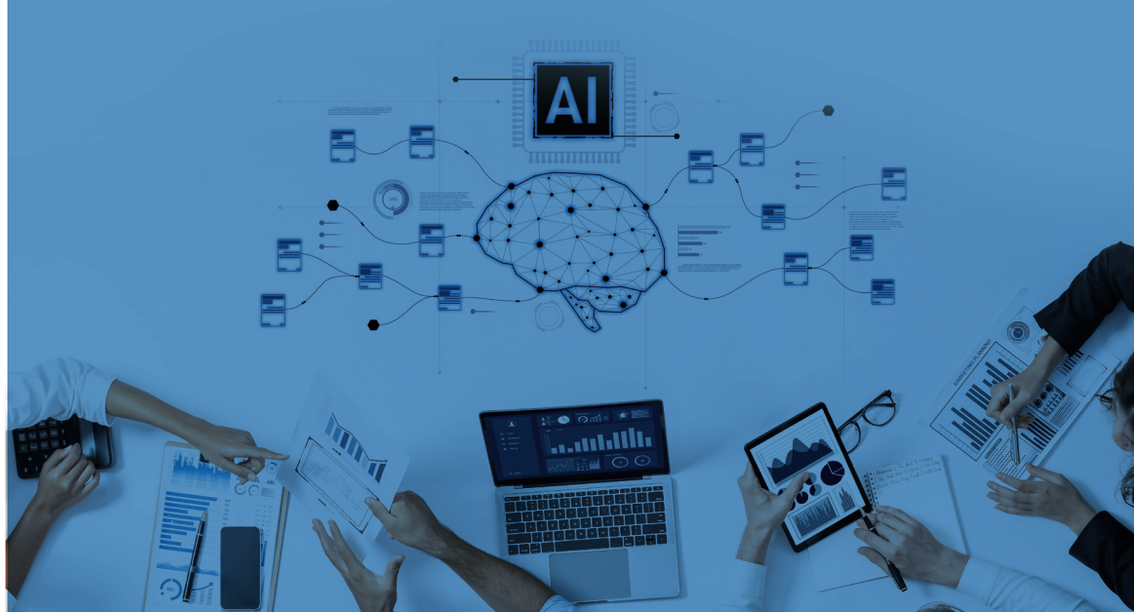
*** 장기 과제: 독립 법률 체계 구축**

- 「AI 책임보험법」 제정: 일정 규모 이상의 사업자에게 배상책임보험 가입 의무화
- 「AI 안전법」 제정: 데이터 보존, 로그 기록, 사고 보고 의무, 안전 설계 기준 등을 규율
- 대규모·중대한 피해에 대해서는 법적 책임이 최종 안전망으로 작동하도록 독립 법 체계 마련
- AI 에이전트 법인격 부여 여부 결정

– 이승철(서울대학교 교수)

II

2025년
기술영향평가 개요



1. 추진 배경

새로운 과학기술이 우리 사회에 미치는 영향은 복잡하고 광범위하므로 이에 대한 선제적 예측과 대응 필요하다. 과학기술의 발전은 경제적 부가가치 창출 등 인류에 긍정적인 효과도 가져다 주지만 환경·윤리 문제 등 예기치 못한 부작용도 초래하기 때문이다.

과학기술은 다양한 분야 및 계층에서 이용되기 때문에 성별 등 사용자 특성에 대한 분석은 해당 기술의 개발 단계에서부터 필요하다. 그래서 사회구성원이 참여한 기술영향평가(TA, Technology Assessment)를 통해 미래 신기술이 초래할 결과를 예측하여 사전 대응 방안을 마련할 필요가 있다.

기술영향평가는 과학기술기본법 제14조(기술영향평가 및 기술수준평가)의 다음 조항을 근거로 추진된다.

① 정부는 새로운 과학기술의 발전이 경제·사회·문화·윤리·환경 등에 미치는 영향을 사전에 평가(이하 "기술영향평가"라 한다)하고 그 결과를 정책에 반영하여야 한다.

② 정부는 과학기술의 발전을 촉진하기 위하여 국가적으로 중요한 핵심 기술에 대한 기술 수준을 평가(이하 "기술수준평가"라 한다)하고 해당 기술 수준의 향상을 위한 시책을 세우고 추진하여야 한다.

③ 정부는 기술영향평가를 실시하는 경우 대상 기술의 성격을 고려하여 성별 등 특성 분석이 반영될 수 있도록 하여야 한다.

④ 기술영향평가와 기술수준평가의 범위 및 절차 등에 관하여 필요한 사항은 대통령령으로 정한다.

2. 평가 내용

본 기술영향평가는 과학기술기본법 제14조에 따라 새로운 과학기술의 발전이 경제·사회·문화·윤리·환경 등에 미치는 영향을 사전에 평가하는 것이다. 「과학기술기본법」 시행령 제23조 제5항에 따라 민간 전문가 및 시민 등 참여가 확대되었는데, 긍정적 영향은 극대화하고, 부정적 영향은 최소화하기 위한 정책적 고려 사항을 이끌어내 정책 수립부터 바람직한 발전 방향을 유도하는 것이 본 평가의 목표이다.

관계 부처는 기술영향평가 결과를 향후 소관 분야 국가 연구 개발 사업의 연구 기획에 반영하거나 부정적 영향을 최소화하기 위한 대책을 추진하는 데 활용할 것이다(과학기술기본법 시행령 제23조 제7항). 과학기술기본법 시행령 제23조 제2항에 따른 기술영향평가의 주요 내용은 다음과 같다.

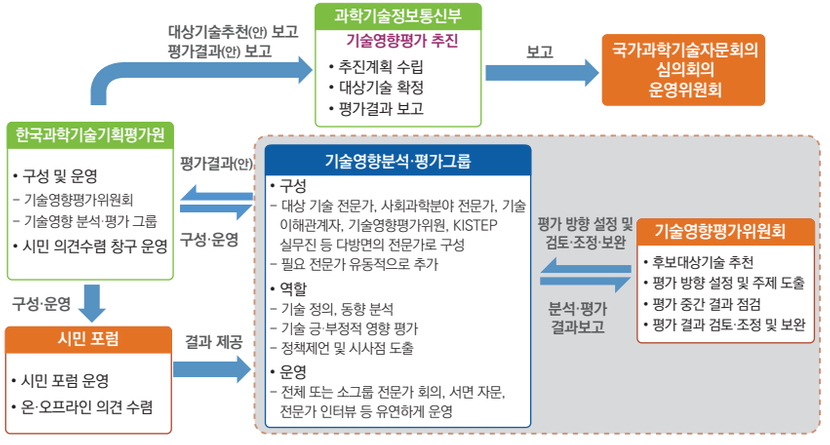
- ㅇ 국민 생활의 편익 증진 및 관련 산업 발전에 미치는 영향
- ㅇ 새로운 과학기술이 경제·사회·문화·윤리·환경에 미치는 영향
- ㅇ 부작용 초래 가능성이 있는 경우 이를 방지할 수 있는 방안
- ㅇ 기술의 성격과 파급 효과가 성별 등 특성에 미치는 영향

3. 추진 체계와 평가 절차

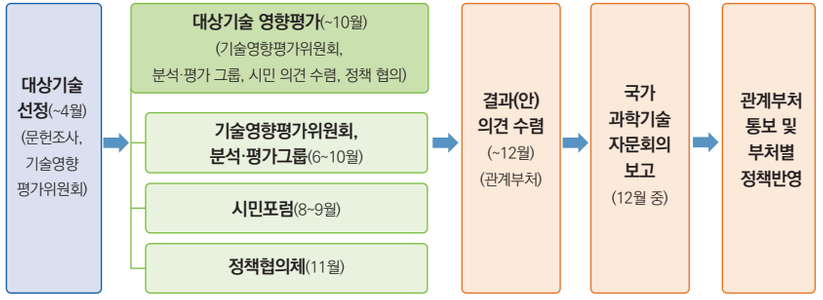
기술영향평가의 총괄 추진은 과학기술정보통신부가 맡는다. 기술영향평가위원회는 대상 기술을 추천하고 평가한 후 방향을 설정하거나 주제를 이끌어낸다. 또 평가 결과를 검토·조정하고 필요한 경우 보완한다.

분석 평가 그룹은 기술 정의 및 동향 분석, 기술의 긍정적·부정적 영향을 평가하며 이슈 및 대응 과제를 제시한다. 시민포럼은 시민 관점에서 이슈 및 대응 과제 제시하며 정책 협의체는 영향 평가 대상 기술 관련 사업·제도 소관 부서·기관·정책 전문가 등 참여하여 수용성 있는 정책 대안 마련한다. 또 한국과학기술기획평가원은 과학기술기본법에 따라 과학기술정보통신부에서 위탁받아 기술영향평가 업무를 수행한다.

〈기술영향평가 추진 체계〉



〈평가 절차〉



시기별 평가 절차는 다음과 같다.

- 기술영향평가 대상 기술 선정(2025년 3~4월)

다음과 같이 미래의 신기술 및 기술적·경제적·사회적 영향과 파급 효과 등이 큰 기술을 대상 기술로 선정한다.

① 현재 등장하고 있는 신기술로서 가까운 미래의 국민 생활에 파급 효과가 클 것으로 예상되는 기술

② 사회적 관심도가 높아 기술영향평가의 유용성과 시의 적절성이 높을 것으로 기대되는 기술

③ 개념과 범주가 명확한 중간 규모(과학 기술 표준 분류상의 중분류 또는 소분류 수준으로 과제 수준 기술은 제외)의 기술

④ 과거에 평가를 실시한 기술 중에서도 환경 변화 및 기술 발전 추이를 고려해 재검토가 필요하다고 인정되는 기술

이외에도 외국 영향 평가, 유망 기술, 전문가나 시민 제안 등 반영하고, 정책적 시의성 및 파급 효과 등을 고려하여 13개 후보 기술(안) 도출한다. 기술영향평가위원회 논의를 통해 특성, 파급 효과의 크기, 사회적 합의의 필요성, 결과의 활용 가능성을 평가하여 5개 기술이 추천되었다. 5개 기술은 AI 에이전트, 소형 모듈 원자로, 도심 항공 교통, 허위 정보 진단, 다기능 로봇이다. 이들 추천된 5개 기술 중 'AI 에이전트' 기술이 대상 기술이 최종 결정되었다.

2025년 기술영향평가 대상 기술로 'AI 에이전트' 기술이 선정된 이유는 첫째, 'AI 에이전트'는 스스로 목표 설정하고 실행하는 자율적 인공지능으로 중요도가 확대되었고 최근 글로벌 기업들은 AI 에이전트를 업무 자동화, 연구 개발, 서비스 운영 등에 적용하며 인간-기계 협업의 새로운 패러다임 형성하고 있기 때문이다. 또 AI 에이전트로 유발되는 사회·경제 전반의 혁신과 함께 윤리·책임·안전 등 예상되는 이슈에 대한 영향 평가 필요하기 때문이다.

- 기술영향평가 대상 기술 예비 분석(2025.5월)

예비 분석을 통하여 기술 특성과 현재 기술 개발·활용 동향 등을 검토하고, 주요 평가 이슈를 이끌어낸다.

- 전문가 분석(2025년 6~10월)

기술전문가, 사회과학, 산업계 등 다양한 분야 전문가 총 10명으로 구성된 분석 평가 그룹은 'AI 에이전트' 기술에 대한 정의와 평가 범위 등을 확정하고, 전문가 관점에서 파급 효과 및 정책적 고려 사항 이끌어낸다.

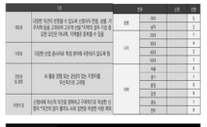
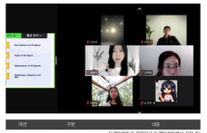
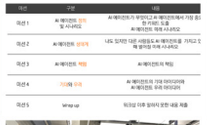
- 시민포럼 제안(2025년 9~10월)

시민포럼은 일반 시민의 관점에서 기술에 관한 생각을 표현하고 기술의 파급 효과, 대응을 위한 권고 사항 등 제시한다. 시민포럼은 성별, 연령, 전공 등을 균형 있게 배분하고 역량 등을 고려하여 선발된 총 8명으로 구성

되었다.

시민포럼은 포럼의 독립성과 시민의 몰입을 강화하여 시민 관점의 중요 이슈 및 시민이 원하는 정책 대안의 발굴을 목표로 한다. 토론 과정에서 참여 시민들에게 미션을 부여하고 그룹 토론 방식을 적용한다. 시민포럼 참여자들이 단계별 5회 미션을 통해 몰입도를 높이고 워크숍에서 관련 내용에 대한 토론을 수행한다. 참가자들에게 적절한 교육을 제공하고 토론을 구조화하여 양극화 및 집단사고의 문제를 완화하고 정보에 입각한 균형 잡힌 결과를 이끌어낸다.

〈시민포럼 절차 및 내용〉

01 프로그램 설계 & 리크루팅		– 총 46명 신청 – 성별, 연령, 지역, 직업 등의 대표성과 다양성, 심화질문에 대한 성실도를 평가하여 8명의 참여자 선발
02 오리엔테이션		– 기술영향평가 개요, AI 에이전트 강연 – 기술영향평가의 의미를 알리고 참여도를 높이며 공동 목표를 설정
03 미션 수행		– 일반시민들이 AI 에이전트 기술 영향의 맥락을 이해하고 몰입할 수 있도록 미션을 부여 – 1개월간 총 5개의 미션을 부여
04 현장 방문		– 국내 AI 관련 대기업 네이버 현장방문 – AI가 실제 산업에 어떻게 적용되고 있는지 살펴보고 우리 생활에 미칠 영향을 예상
05 시민참여 포럼		– 참여자들이 숙고를 바탕으로 다양한 관점을 이야기할 수 있도록 워크샵 진행 – AI 에이전트가 우리 삶에 미치는 긍정/부정적 영향을 토의하고 필요한 정책 및 규제 아이디어 도출
06 결론 도출		– 오리엔테이션, 5회 미션, 시민참여 포럼에서 논의된 내용을 바탕으로 일반 시민들이 생각하는 AI 에이전트의 기대와 우려, 정책 제안 정리

- 정책 협의체 논의(2025년 10~11월)

정책 협의체는 기술영향평가에서 제안된 정책 수요의 수용성 향상을 위해 관련 부처와 전문가로 구성·운영되며 정책을 논의한다.

- 기술영향평가 결과(안) 마련(2025년 11월)

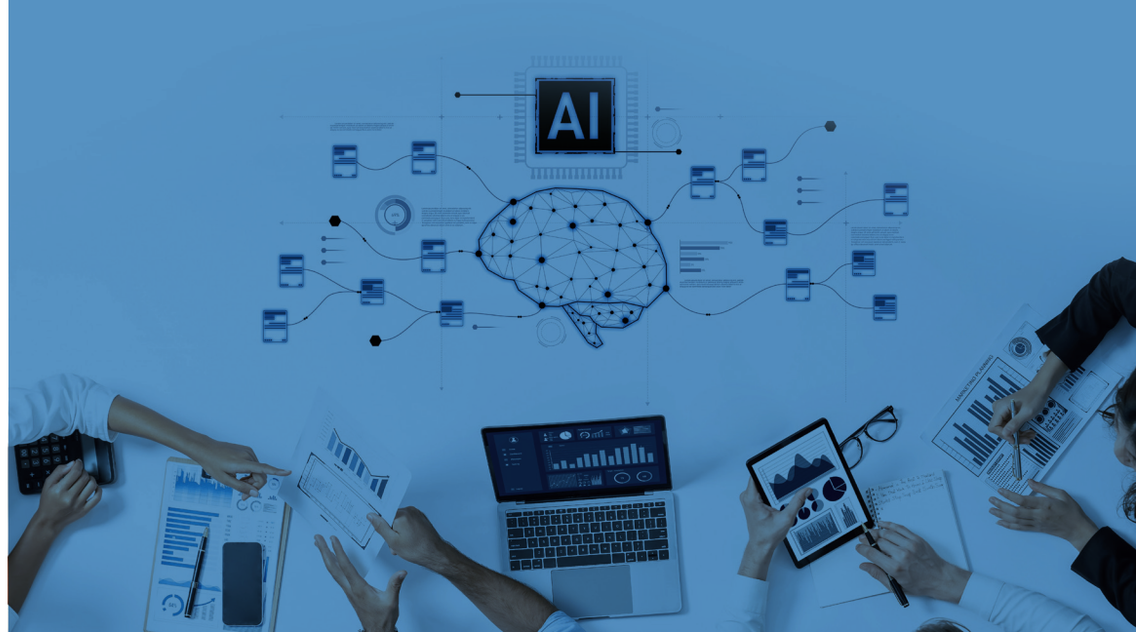
전문가 및 시민포럼에서 제안된 주제별 이슈, 긍정적/부정적 파급 효과 등을 토대로 이슈 대응을 위한 정책 수요를 찾아낸다. 이 정책 수요를 기반으로 전문가 설문을 통해 주요 정책을 이끌어내고 정책의 중요성, 시급성, 효과성, 실현 가능성 등을 분석한다. 이를 바탕으로 결과(안)을 마련하여 관계 부처에 의견 회람을 한다(12월).

- 국가과학기술자문회의 결과 보고(2025년 12월)

과학기술기본법 및 국가과학기술자문회의법에 따라 결과 보고를 한다.

III

2025년 기술영향평가 결과 : AI 에이전트



1. 기술 개요 및 평가 범위

AI 에이전트(AI Agent) 기술의 정의

AI 에이전트는 인간처럼 목표를 스스로 세분화하고, 외부 자원을 활용해 다단계 작업을 수행하는 자율적 AI 시스템이다. 최종 목표를 달성하기 위한 세부 과정(Sub-task)을 스스로 잘게 나누고 정의하며 작업을 수행하는 자율적 AI 시스템을 포함한다. 본 기술영향평가의 'AI 에이전트'는 관련 개념인 에이전틱 AI(Agentic AI), 물리적 AI(Physical AI)를 모두 포괄하는 개념으로서 조작적으로 정의한다.

AI 에이전트 및 관련 개념은 다음과 같이 나눠서 설명할 수 있다. AI 에이전트는 인공지능을 통해 특정 작업을 자동화하는 단일 시스템이고 개별적 존재로 특정 작업에 특화되어 정해진 범위 내에서 임무를 수행할 수 있다. 에이전틱 AI는 여러 AI 에이전트가 협력하여 복잡한 문제를 해결하는 복합 시스템이다. 다중 에이전트 또는 고도화된 자율 AI로, 포괄적인 목표 달성을 위해 여러 기능을 통합하고 복잡한 의사 결정을 자율 처리한다. 물리적 AI는 물리적 환경에서 작동하는 AI 에이전트 시스템으로, 로봇 등 물리 에이전트에 국한되며 물리 환경과 상호 작용을 위한 제어 기능을 통합한다.

〈AI 에이전트 및 에이전틱 AI 비교〉

구분	AI 에이전트	에이전틱 AI
범위	제한된 작업 자동화	여러 AI 에이전트 협력, 복잡한 문제 해결
자율성	제한적 자율성	높은 자율성
목표 설정	주로 외부에서 설정	자체적으로 목표 설정 및 계획 수립
복잡성	단순 작업 자동화	복잡한 문제 해결
예시	챗봇, 개인 비서	자율 주행, 지능형 자동화
	특정작업 수행하는 숙련된 전문가(도우미)	여러 전문가들이 협력하여 복잡한 프로젝트를 수행하는 팀

Ranjan Sapkota et. al. (2025), AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges

기술의 평가 범위[1]

AI 에이전트의 핵심 기술과 상호 작용에 대한 두 개의 축으로 구성하여 기술영향평가를 수행한다. 핵심 기술은 AI 에이전트 기술의 핵심 요소인 추론, 행동과 함께 자율성을 고려한다.

〈AI 에이전트의 핵심 기술 구성〉

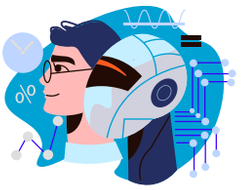
핵심 기술	내 용
추론 (Reasoning)	○ (정의) AI가 주어진 정보로부터 의미를 도출하고, 목적 지향적 판단을 수행하는 능력 ○ (예시) 복잡한 문제 해결을 위한 멀티스텝 추론 및 계획, 명시되지 않은 정보의 추정 및 상황 판단
행동 (Action)	○ (정의) 환경이나 시스템에 실제로 영향을 주는 실행력 ○ (예시) 로봇의 물리적 조작, 소프트웨어 에이전트의 자동화된 업무 처리
자율성 (Autonomy)	○ (정의) 인간의 개입 없이 목표를 스스로 설정하고 달성하는 능력 ○ (예시) ChatGPT 기반 에이전트가 스스로 정보 탐색, 응답 작성, 반복 수행

AI 에이전트가 상호 작용하는 대상에 따라 A2H, A2A, A2E로 나눌 수 있다(본 논의를 위해 자체적으로 구분 정의).

〈AI 에이전트의 상호 작용 구분〉

상호작용	내 용
AI 에이전트-인간(A2H)	○ (정의) 인간과 AI 에이전트 간의 정보 교환, 의사소통, 협력 관계를 의미 ○ (예시) 디지털 비서, 헬스케어 상담 에이전트, 교육용 AI 튜터
AI 에이전트-AI 에이전트(A2A)	○ (정의) 복수의 AI 에이전트가 정보를 공유하거나 협력하여 공동 목표를 달성하는 관계 ○ (예시) 자율주행차 간 협력 운행, 에이전트 기반 물류 최적화 시스템, AI agent orchestration
AI 에이전트-환경(A2E)	○ (정의) AI 에이전트가 물리적 또는 디지털 환경으로부터 데이터를 수집하고, 이에 적응하며, 반응하는 능력 ○ (예시) 로봇의 물리적 이동, 스마트팩토리 내 환경 변화 감지 및 대응, 게임 환경에서 학습하는 AI 에이전트

2. 기술영향평가 결과



(1) 평가 주제

예비 분석을 통해 AI 에이전트 관련 주제 및 이슈를 이끌어낸 후, 분석·평가 그룹에서 나눈 논의 및 토론을 바탕으로 평가 주제를 선정한다. 각 주제 및 이슈는 경제·사회·문화·윤리 등 다양한 분야에 복합적인 영향을 주며, 이를 고려하여 주제 단위의 의견서 작성하였다.

〈기술영향평가 주제 및 주요 이슈〉

주 제	주요 이슈
① 융합 및 혁신	타 분야와 융합
	신산업 창출
	기업 경영(협업인프라 및 시범적용)
② AI 에이전트와의 관계	AI 에이전트와의 정서적 관계
	상호작용
	기술 종속적 문화의 확장
③ 노동 변화	노동 대체와 변화
	재교육
④ 개인정보보호/활용	자기결정권 보장
	리터러시
	데이터 추적 및 검증
⑤ 사회적 수용성과 미래 교육	사회적 수용성
	접근성 격차
	수용 특성

주 제	주요 이슈
	세대 수용성
	미래 교육
	특성 평가
⑥ 책임	책임소재의 불명확성
	대응 방안
⑦ 윤리	윤리 정의
	가치 충돌
⑧ 주권 및 안보	글로벌 대응
	군비 경쟁과 안보
	관련 산업 종속
⑨ AI의 지속 가능성	에너지
	지속 가능성
⑩ 시민사회와 민주주의	시민 참여와 정치
	여론과 정보 다양성

(2) 주제별 현황

융합 및 혁신

다른 기종 시스템(ERP·CRM 등) 사이를 툴 호출·API 어댑터·이벤트 버스로 연결, 부서·기관 간 데이터나 업무 흐름을 연동시킨다. AI 멀티에이전트 활용으로 '단일 기능'에서 '연결형 번들'로 전환된다. 오케스트레이션(Orchestration) 기능으로 단순 병렬이 아닌, 목적 지향적인 워크플로우 구성 및 신뢰 가능한 에이전트 생태계 구성이 가능해지는 것이다.

구글·오픈AI 등 글로벌 기업이 AI 에이전트 시장에서 앞서 나가고 있으나, 스타트업의 오픈소스를 토대로 한 신산업 창출 경쟁이 치열하다. 지난 2년간 AI 에이전트 관련 스타트업에 유입된 투자 자본은 20억$(약 2.78조원)를 초과했다(출처: 딜로이트 2025 첨단기술·미디어·통신산업 전망).

2025 첨단 기술·미디어·통신 산업 전망을 관측한 딜로이트 보고서(2025년 1월 23일)에 의하면 2025년 생성형 AI를 도입한 기업 중 25%가 AI 에이전트를 시범 도입하거나 기술 검증에 나설 것으로 전망된다. 이후 2027년에 이르면 그 비율은 50%로 늘어날 것으로 예상된다. 스타트업과 기존 테크 기업들이 수익 잠재력을 기대하고 AI 에이전트 개발에 힘을 쏟고 있으며 지난 2년간 AI 에이전트 개발 스타트업에 유입된 투자 자본은 20억 달러가 넘는다.

또한 빅테크 기업들이 AI 에이전트 플랫폼과 상호 운용하여 프로토콜을 출시했고 에이전트 간 협업을 위한 기술적 인프라 구축했다. 앤트로픽(Anthropic)의 모델 콘텍트 프로토콜(Model Context Protocol, MCP)은 AI 에이전트와 외부 시스템을 표준화된 방식으로 연결하고 있다. 에이전트 협업 인프라를 기반으로 서로 다른 전문성을 가진 멀티에이전트 시스템이 의료 분야를 중심으로 연구 및 시범 적용되고 있다. 대표적으로 XR 레포트젠(XRReportGen)의 흉부 X-ray 보고서 생성, 메드이미지팔스(MedImageParse)의 아홉 가지 이미징 모달리티 분석, 메드이미지인사이트(MedImageInsight)의 유사 사례 검색 등을 들 수 있다.

AI 에이전트의 상호 작용에 대비한 체계를 준비하고 관련 산업 촉진과 규제를 마련할 필요가 있다. 산업 간 에이전트 네트워크와 표준 프로토콜 구축 필요성을 강조하고 의료, 보험, 유통 등 분야 간 협업 체계 마련해야 한다. 또한 규제 프레임워크 구축을 통해 책임 부여 및 자율성 범위를 보장하고 규제 샌드박스 적용 및 실증 사업을 추진할 필요가 있다. 에이전트 자율성 보장과 개인정보보호법, 의료법 등 관련 법령의 유연한 적용을 통한 실증 환경 구축도 필요하다.

AI 에이전트와의 관계

AI 에이전트는 사용자 대리 과정에서 사용자의 개인 정보와 인간의 특성을 파악함으로써 정서적 교감과 상담 기능을 제공한다. 사회적 유대 강화와 심리적 안정감을 부여함으로써 기존 복지 체계의 사각지대를 보완할 수 있다. 미국 10대 다수가 AI 동반자(companion)를 사용해 본 경험이 있으며, 정기 이용 비중도 높다고 보고되고 있다.[2]

다수의 AI 에이전트가 상호 협력·분업하여 복잡한 문제를 해결함으로써 효율성 제고하고 있다. 물류, 에너지 관리, 교통 관리, 스마트시티 운영 등에서 집단적 문제 해결도 가능하다. 하지만 자율성이 부여된 상호 작용의 확대에 따른 에이전트 통제 불능 우려도 있다. AI 에이전트 간 경쟁적 상호 작용에서 사기, 권한 오남용의 위험이 있으며 인간-AI 간 대리인 딜레마와 책임 소재가 불분명해지는 문제가 발생할 수 있다. AI 에이전트가 특정인의 말투와 습관을 모방하여 지인을 사칭하는 방식으로 금융 사기에 악용될 가능성도 매우 높다.[3]

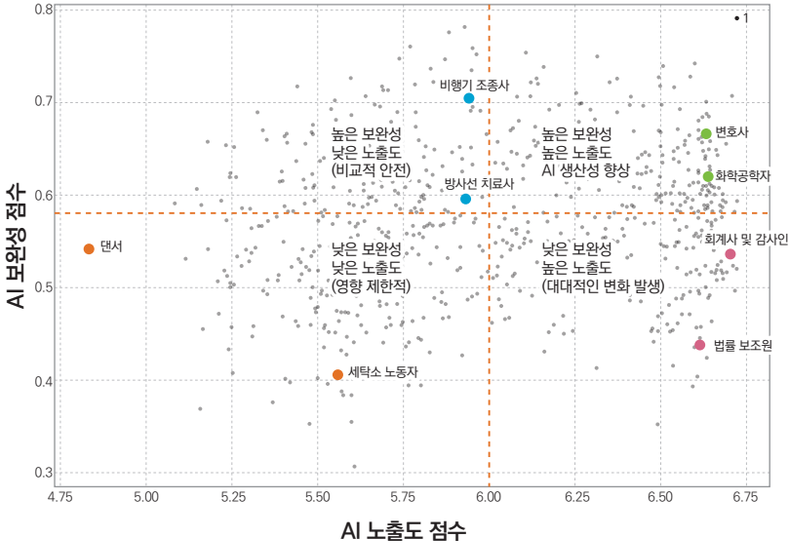
과도한 정보 의존으로 주객이 전도되어 인간의 자율성과 비판적 사고력이 약화될 수 있다. 기술적으로 종속될 우려가 있는 것이다. AI 에이전트에 의한 과도한 최적화로 경험 영역이 축소되는 '인지적 갇힘 현상(echo chamber)' 발생 가능성도 있다. 알고리즘 최적화 및 추천 시스템 의존이 편향성을 심화시키고 사회적 분열을 확대할 수도 있다.

인간과 AI 에이전트 간 정서적 상호 작용 가능성과 위험성에 대해 나눈 사례별 정책 논의 결과는 다음과 같다. 우선 노년층이나 일반인의 정서 동반자, 디지털 치료제 연계에 따른 심리 치료 등 목적에 따라 AI 에이전트의 역할을 구분하여 R&D 및 제도 추진해야 한다. AI에 대한 과도한 의존과 주체성 약화를 방지하는 제도적 장치가 필요하고 AI 에이전트 간 상호 작용에서 예상되는 경쟁과 통제 불능 상황에 대비하여 법, 제도, 가이드 라인 등을 구축해야 한다. 현재 AI 상호 작용의 패턴, 기법 등이 명확히 규정되지 않으므로 추가적인 기술 개발 및 관찰도 필요하다.

노동 변화

많은 직종에서 일하는 방식의 변화와 생산성 향상을 이끌 것으로 예견되며 이에 따른 노동 대체가 예상된다. 기술 혁신은 늘 노동시장의 변화를 이끌어 왔으나, AI 에이전트는 전문직을 포함한 광범위한 직업군에 노동 변화를 초래할 것으로 보인다. 조지타운 대학의 연구에 따르면 회계, 감사, 법무 등의 전문직은 AI 노출도가 높으면서 AI 보완성은 낮아 가장 취약한 직업군이다.

〈직업군별 AI 영향 정도〉



출처 : 한국산업기술진흥원, 『인공지능(AI) 시대 인력 개발의 미래』, p.3

생성형 AI 에이전트의 도입으로 다수 직군에서 근로자가 지닌 기존 기술의 노후화가 급격하게 진행되므로 재교육 필요하다. AI 에이전트의 도입으로 언어, 전산, 자료 수집 및 분석 등 근로자가 축적한 기존의 기술이 빠른 속도로 노후화되고 대체될 것으로 예상된다.

저숙련 일자리뿐만 아니라 고숙련자 계층까지 AI 에이전트에 의해 대체될 우려가 있으며 대응 방안이 필요하다. 기존 노동자에 대한 재교육을 지원하는 방안 중심의 정책뿐만 아니라 새로운 고숙련자 육성 및 숙련 진입 단계에 지원하는 정책을 추진해야 한다. 기본 소득, 신산업 창출을 통한 고용 대체 등 새로운 거시적 노동 정책 연계도 필요하다.

개인 정보 보호와 활용

개인 정보 활용은 점차 확대될 것으로 예상되며 개인의 잊혀질 권리, 개인 정보 미활용 등의 자기결정권 수요가 증가할 것으로 예상된다. 기존 개인 정보 활용 동의는 익명화를 거치고 제공하는 경우가 많지만, 점차 제공하는 데이터가 많아지면서 재식별 가능성이 생긴다. 넓은 의미에서의 개인 정보 보호를 적극적으로 수행하기 위해서는 데이터를 활용하는 시점에 데이터 소유자의 의사를 확인하는 과정이 필요하다.

개인 정보는 한번 유출되면 돌이킬 수 없으므로 관리 및 부작용에 대한 충분한 교육이 필요하다. 개인 의견을 과시하고 새로운 사실에 대한 다양한 해석과 의견을 제시하는 등의 사회적 관계망을 활용하고 있으나, 데이터의 재사용은 제어가 곤란한 상태이다.

리터러시 시행으로 개인 정보 데이터가 어디까지 사용될 수 있고, 데이터의 가치나 악의적 활용에 대해 충분히 교육할 필요가 있다. 이와 관련된 교육은 청소년기에 진행되어야 하고, 자신을 알리는 행위와 개인 정보를 보호하는 행위 사이의 의미에 대한 교육이 필요하다.

개인 정보를 넓은 의미로 해석하고 이를 적법하게 활용하기 위해서 자신의 데이터를 주관적으로 관리하고 제어할 필요가 있다. 데이터는 개인의 행동이나 생각을 포함할 수 있으며 이 경우, 모두 개인 정보로 볼 수 있기 때문에 개인 정보에 대한 단계별 정의도 필요하다. 세부적인 데이터의 정의에 따라서 '매우 민감' 혹은 '민감'한 개인 정보, 생활 데이터, 사고 데이터 등으로 구분하여 저장하고 제어하는 기술 연구도 있어야 한다.

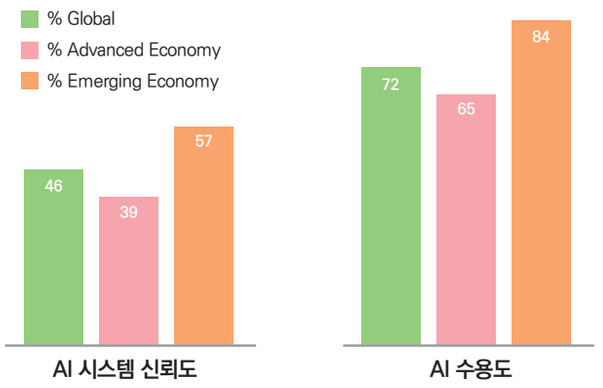
AI 에이전트 활용 시 발생할 수 있는 개인 정보 남용과 교환에 대응하는 정책 마련이 필요하다. AI 에이전트는 정의되지 않은 방식으로 개인 정보를 활용할 가능성이 높아, 블랙박스, 로그 추적 등 이에 대한 인증이나 기록을 제도화할 것을 제안한다.

개인 정보 제공 및 보호에 대한 '자기결정권'을 보장해야 하고, 에이전트 간 정보 교환 시 기준을 마련해야 한다. 메모리 장기 기억으로 인한 'AI 에이전트 개성' 문제에 대응을 위해 제도적 통제 방안이 필요하고 마이데이터 등 개인 정보 통제 체계, 데이터 패스포트 등을 활용하는 촉진 정책 방안이 제시되어야 한다.

사회적 수용성과 미래 교육

AI 시스템 신뢰도는 지역별·계층별로 편차가 크다. 선진국에서는 AI 신뢰도가 39%에 불과하지만 신흥국은 57%에 달하는 상황이다. 이는 AI 수용도에도 영향을 미쳐 신흥국은 84%, 선진국은 65%의 AI 수용도를 확인할 수 있다.

〈AI 시스템 신뢰와 수용도(KPMG)〉



https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html

AI는 개인의 능력·접근성·학습 기회·경제력에 따라 혜택의 격차가 커져 오히려 기존 사회·경제적 격차를 심화할 수 있다. 도시 고학력층 위주로 AI 도구와 데이터 활용이 확산되고 취약 계층은 활용이 저조하여 학습 기회와 정보 접근에 격차가 확대될 가능성이 있다.

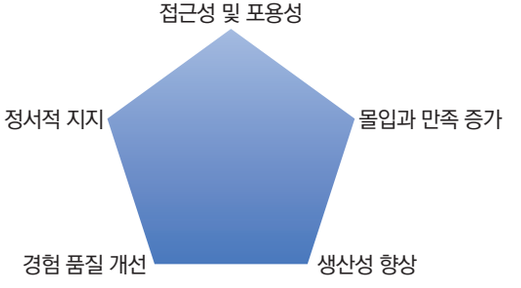
교육 분야에서 교사의 AI 활용 역량과 지원 수준 차이로 학교별 도입 편차가 발생하고, 일부 교사들은 두려움 때문에 AI 활용을 기피하는 현상도 보이고 있다. 미국에서 교외의 부유한 학군이 도시·농촌의 저소득 학군보다 AI 활용에서 앞서 있다는 초기 징후가 포착되었다.[4]

AI 에이전트가 보편화된 세대와 그렇지 못한 세대 간 인식과 활용에 차이가 생겨 갈등이 심화될 수 있다. AI 의존 세대는 높은 멀티 태스킹 능력을 보이는 동시에 낮은 집중력과 높은 의존성으로 전통적 교육 성과가 저하될 가능성도 있다.

미래에는 AI 에이전트 기술에 대한 사회적 수용성이 비교적 높아질 것임에도 불구하고 '탈숙련'이나 사용자 주체성 문제 등을 해결할 필요가 생길 것이다. 사회적 수준에서 특정 전문성을 배울 기회나 동기가 사라져서 궁극적으로는 사회 전체가 전문 역량의 '탈숙련' 위험에 빠지게 될 수도 있다. 이에 AI와 성공적으로 협업하면서도 특정 분야의 전문가로서의 '역량'을 유지할 수 있는 방안을 찾는 것이 필요하다.

AI 에이전트는 기술적 성능에 더해 사용자가 가진 사회적·심리적 특성에 의해 수용 여부가 크게 달라지는 대표적 기술이다. 사용자가 AI 에이전트를 기술이 아닌 사람처럼 대하는 경향이 반복적으로 보고되며 상호 작용에 따라 복잡한 반응 양상도 예상된다. 성별, 연령, 지위, 장애, 문화 등 이용자 특성에 의해 접근성·정확도·위험이 비대칭으로 발생하고 데이터 편향이 사회적 불평등을 재생산할 수도 있다.

〈이용자 특성과 AI 에이전트 가치 지도〉



특정 분야에 국한되지 않고 수용성 제고를 위한 정보 신뢰도가 확보되어야 하고 미래 교육의 중요성이 강조되어야 한다. AI 에이전트의 오용으로 인한 사회적 부작용 방지를 위한 교육과 윤리 강화도 필요하다. 시민 리터러시 교육, 에이전트 사용자의 자격인증제도, 공공 감시나 옴부즈만 등 피해 구제 시스템을 제안한다.

책임

자율적으로 행동하는 AI 에이전트의 특성상, 그 결과 발생하는 행위와 영향에 대한 책임 귀속 문제가 불가피하게 제기된다.[5] 현행 법 제도는 주로 인간 중심의 법적 책임 관념에 기반해 작동하므로, AI 에이전트의 행위 결과에 대한 책임 주체는 불명확한 상태이다. 특히 의료·군사·금융 등 고위험 분야에서 사고 발생 시 책임 문제는 더욱 큰 논란 대상이 될 수 있다. 책임 소재의 불명확성은 사회적 신뢰와 안전성에 직접적 위협이 될 소지가 있음은 다음 사례에서 찾아볼 수 있다.

- Raine v. 오픈 AI(2025년) : 16세 청소년의 자살 사건 관련, 부모가 오픈 AI 상대로 "자살 조장·방치" 소송 제기[6]
- 에어 캐나다 챗봇 사건(2022년, 2024년 판결): AI 챗봇의 잘못된 환급 정보 제공에 대해, 법원이 회사 책임 인정, 보상 명령[7]
- 오퍼레이터 사건(2025년): 오픈 AI의 '오퍼레이터' 에이전트가 승인 없이 장보기 주문 실행. 사전에 구매 확정 절차를 거친다는 안전 장치가 있었음에도 이를 우회[8]
- 테슬라 자율 주행 사고(2019, 2025 판결): 자율 주행 모드에서 발생한 사망 사고에 대해 미국 플로리다주 배심이 운전자 과실과 제조사 책임을 병행 인정[9]

이에 법적, 윤리적 책임 소재의 불명확성 해소가 필요하며 법·제도 준비 및 사회적 대응 방안 마련도 있어야 한다.

정책적으로는, AI 에이전트가 자율적으로 수행한 행위에 대한 법적 책임이 논의되어야 한다. 통제 가능한 범위에서의 면책 여부, 손해배상 책임 주체에 대한 규명이 필요하고 기업의 서비스 운영 책임, 개발자의 간접 책임, 사용자 책임 등 다양한 계층별 책임 소재 재정립을 추진해야 한다. AI 에이전트 행동 이력을 기반하여 책임 추적이 가능하도록 기술 및 제도적 장치가 마련되어야 하고 사고에 따른 배상·보험제도 개편과 감시, 개입, 조정 등을 수행할 전담 기구를 설치할 필요도 있다.

윤리

AI 윤리는 사회적 윤리, 집단적 윤리, 개인적 윤리 기준이 있으며, 이들은 상호 간에 충돌하면서 지속적으로 성장해왔다. AI의 윤리적 관점이 사회적 이슈로 대두될 것이며 이를 적극적으로 반영하는 방법에 대한 논의가 시작되고 있다. AI의 큰 걸림돌 중의 하나로 윤리적 판단을 할 수 있는가이다. 사회적 가치 변화의 AI 학습은 긍정적인 사항으로 등장하고 있지만 사회적 가치를 평가하는 것은 매우 어렵다. 그래도 가치 판단의 근거를 제시하는 것은 사회적 합의를 얻는 방법임은 분명하다. 변화하는 사회의 가치를 반영하고 기존의 가치를 변화하는 방법으로 지속적으로 사회에 공유하는 활동은 사회 통합의 차원에서 긍정적인 활동으로 볼 수 있다. AI의 윤리적 판단은 논쟁의 중심이 될 수 있으며 사회적 변혁의 시기에는 사회적 갈등과 혼란을 불러 일으킬 수 있다.

사회적 가치 가치 사이의 충돌이 있는 경우 사람이나 AI의 윤리 가치 판단을 요구한다. AI의 윤리적 가치 충돌은 개인보다는 집단에 의해서 결정되고 권력이 큰 집단에 의해서 결정될 가능성이 있다. 사회의 여론을 형성하는 과정에서 AI 윤리에 대한 다수의 침묵은 일부에 의해서 주도될 수 있어 사회적 부작용이 우려된다. 사회적 문제가 발생하는 경우 한쪽의 일방적인 손해보다는 상호 간의 가치 판단을 공유하고 논의하는 노력이 필요하다.

정책적으로, 데이터 활용 가이드 라인, 윤리 간 충돌 시 해결 방안, 사회적 책무와 윤리적 부작용 대응을 위한 리터러시 교육 등을 검토할 필요가 있다. 저작권, 라이선스, 데이터 수집의 정당성, 타인의 권리 침해 등 법적 쟁점을 포함하여 정책 제안을 할 만하다. 법적 쟁점으로는 크롤링 AI 활용의 순환적 피해(무단 활용→보호 요청→우회적 침해 반복) 등을 예로 들 수 있다. 다양한 분야가 포함되므로 다분야 전문가 참여를 통한 다학제적 접근 및 연구가 필요하다.

주권 및 안보

미국과 중국의 AI 패권 경쟁이 가속화되고, EU는 AI Act 제정 등 규범 정립을 주도하며 각국 AI 주도권 확보에 총력을 기울이고 있다. 미국은 기술 주도권을 유지하고 글로벌 경쟁 구도에서 우위를 확보하기 위한 전략적 대응으로 국가 AI 전략을 마련하고 있다. 이 계획의 원제는 「America's AI Action Plan」 / 「미국 AI 실행계획」이다.

중국은 국제 규범 설정 역량을 강화하고, AI 외교를 통해 글로벌 영향력을 확대하기 위해 글로벌 AI 전략을 마련했다. 이 계획의 원제는 「人工智能全球治理行动计划」 / 「글로벌 AI 거버넌스 실행계획」이다.

AI 기술이 무기 자동화, 드론 정찰, 사이버전 등에 활용되며 현대전 양상을 변화시키고 있어 안보 전략에 AI 통합을 추진할 필요가 대두하고 있다. 군사적 활용이 방어 역량을 혁신적으로 향상시킬 수 있지만 예측 불가능한 자율 시스템 간 상호 작용과 새로운 형태의 안보 위협을 초래할 가능성도 있다. 자율 무기의 오작동이나 오남용, AI 판단 오류로 인한 우발적 충돌 가능성 등 새로운 안보 위협과 윤리적 딜레마가 대두되고 있다.

핵심 기술과 데이터 인프라가 일부 국가 및 거대 기업에 집중되면서 외국 기술 의존에 따른 기술 주권 약화도 우려된다. 자체 AI 모델이나 인프라가 없을 때 외국 기업이 서비스를 중단하거나 통제에 대응할 수 없어 'AI 식민지'로 전락할 위험도 있다. AI 에이전트 기술이 공공 행정에 실험적으로 도입되기 시작하면서 정책 집행의 효율성과 주권적 의사 결정 능력이 동시에 주목 대상이 되고 있다.

AI 에이전트 기술이 외국 플랫폼 및 모델에 종속될 경우, 기술 및 데이터 주권이 침해될 우려가 있다. 미국과 중국 AI 경쟁에 따라 각각의 생태계 구축 및 편입 압력이 커지는 중이며 우리나라는 하이브리드 전략을 추구하고 있다. 국내 생태계 보호 및 국산 AI 에이전트 육성을 위한 K-agent, K-MCP 등 R&D 및 제도적 전략을 세울 필요가 있다. 군사 안보와 함께 산업 안보 관점에서 정책을 제안해야 한다.

AI의 지속 가능성

LLM 학습·운영 시 많은 전력이 소모되며, AI 에이전트는 지속적 실행을 해야 하므로 상시 전력 소모가 크다. AI의 학습과 추론 시 막대한 데이터와 계산 자원이 필요하며, 이를 위한 데이터센터는 상당량의 전력을 소모해야 한다. 데이터센터의 냉각과 전력 공급(생산) 과정에서 다량의 탄소가 배출되며 전력 소모를 줄일 수 있는 경량 AI 모델 개발이 필요하다. AI 에이전트 기술 개발 과정에서 NPU, TPU와 같은 저전력 AI 칩, 엣지 컴퓨팅, 지능형 자원 관리 등과 같은 기술이 활성화되어야 한다.

지속 가능한 AI 에이전트를 담보하기 위해서는 데이터 수집·활용의 합법성·윤리성 확보가 필요하다. 편향된 데이터 사용은 기술 신뢰성을 떨어뜨리고 사회 지속 가능성에 악영향을 미친다. 대규모 AI 인프라를 보유한 소수 기업이 기술을 독점하게 되면 AI 혜택에 대한 편중과 독과점이 발생

할 수 있다.

정책적으로, 지속적인 AI 에이전트의 서비스 유지 및 제공을 위해 자원과 에너지 사용의 효율성 확보와 사회적 책임 논의가 필요하다. 저전력 소요 알고리즘 개발, 신재생 에너지원 보급 확대 등이 필요하고 AI 에이전트 개발 및 운영 기업의 사회적 책임 강화 및 윤리적 기준 준수가 지속가능성 확보의 핵심이 되어야 한다.

시민사회와 민주주의

AI 에이전트를 통한 시민 참여 플랫폼이 확산되고 있으며 정책 결정에서 시민 의견 반영 필요성 증대되고 있다. 공공 정책 형성 과정에서 AI가 시민의 의견을 매개하고 요약·전달하는 역할을 수행하고 있다. AI의 윤리적 판단 기준에 관한 사회적 합의, 대의제 한계 보완, 기술 불신 해소, 공공 감시 기능 강화를 위해 시민 의견 반영이 필요하다.

여론은, AI 개인 맞춤형 정보를 제공하며 여론의 기저를 형성하는 정보 생태계의 핵심 행위자로 떠오르고 있다. AI 에이전트가 정보 필터링 및 요약, 추천을 통해 여론 형성에 직간접적 영향력을 행사하는 것이다. 기존 언론 및 미디어 플랫폼 중심에서 AI 에이전트 중심의 여론 형성으로 전환하며 시민이 이를 인지하고 능동적으로 대처할 필요가 있다.

정책적으로, AI 에이전트가 정치 참여, 여론 형성, 사회운동 등에서 미치는 영향을 고려하여 대응 방안을 마련해야 한다. 사용자와의 공감 능력 및 정보 전달력으로 인해 특정 사상 또는 정치적 편향을 강화할 가능성도 제기된다. 민주주의 원칙과 표현의 자유를 보장하는 방향의 가이드 라인이 마련되어야 하고 공공 알고리즘 투명성 및 설명 가능성을 확보할 필요가 있다. 선거 운동, 여론 조작, 편향 판단 어려움 등 대응을 위한 정책 제안도 추진되어야 한다.

(3) 주제별 정책 수요

AI에이전트 관련 주제별 현황에서 우리 사회에 큰 영향이 예상되는 긍정적/부정적 파급 효과가 확인되었다. 세대 윤리나 문화의 변화 등 사회 구성원의 생각 변화와 함께 일자리, 산업, 정치 등 사회 구조에 대한 큰 변화가 예상된다. 이에 기술영향평가를 통한 혁신 기술의 시민 담론 형성과 함께, 예측되는 미래 이슈 및 사회 변화에 대한 정책 대응이 필요하다. 적극적인 정책 분석와 환류를 통해 미래 기술의 시너지 확대 및 부작용 감소를 추진해야 한다.

〈주제별(융합 및 혁신) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	규제 샌드박스 운영 (개인정보처리 등 제도 개선 포함)	AI 에이전트와의 관계, 주권 및 안보, 책임(제도실험)
	에이전트 규제 프레임워크 수립	책임, 관계 (자율성 기반 법적 책임 논의)
	에이전트 자율성 보장(에이전트 법인격 시범 부여, 법령 유연 적용)	-
	멀티에이전트 보안 프레임워크 구축	-
② 제도, 행정	국제표준 대응 및 상호운용 생태계 구축	주권 및 안보(K-Agent 프로토콜, 공공 생태계 구축)
	공공분야 AI에이전트 서비스 확대 및 테스트베드 활용	수용성(공공플랫폼), 시민사회(공공 AI 실증)
	에이전트 책임보험 의무화	책임(사고 보장 및 피해 구제)

정책 분류	정책 제안	연관있는 주제(정책)
③ R&D	산업 특화 AI에이전트 서비스 개발 및 실증(특화 파운데이션 모델, XAI 등)	노동, 지속가능성, 수용성 등 (공공 실증 및 민간 실증)
	산업간 에이전트 네트워크 및 표준 프로토콜 구축	-
	멀티에이전트 협업 시나리오 실증	관계 (AI 상호작용 및 정서적 협업 실증)
④ 기타	다양한 산업 특화 모델 개발 및 실증	
	창업팀 발굴 및 글로벌 진출 지원	노동(스타트업 육성)

〈주제별(AI 에이전트와의 관계) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	인간-AI 정서 상호작용 가이드라인 마련	-
	AI 에이전트 통제불능 상황 대비 법제	책임(피해구제, 통제체계)
② 제도, 행정	AI 기업의 옴부즈만 제도	
	AI 에이전트 운용 공인 인증제 도입 (서비스 기업, 기관 대상 담당자 지정)	책임, 시민사회(AI 인증제, 시험제)
③ R&D	정서 교감 및 심리상담 기술 개발	-
	상호작용 기반 자율성 시뮬레이션 기술	-
	AI 관계 변화에 대한 사회문화 영향평가	-
④ 기타	정서 동반자 AI 시범사업 (일반챗봇, 심리치료 등)	수용성(공공 AI 플랫폼), 시민사회(공공 AI)
	AI 자율성 인식 교육 프로그램 개발	시민사회, 수용성 (비판적 사고 및 디지털 리터러시)
	인간 자율성 보호 위한 리터러시 교육	시민사회, 수용성 (비판적 사고 및 디지털 리터러시)

〈주제별(노동 변화) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	사회보험의 포괄적 적용과 개인계정 도입	-
② 제도, 행정	직무 전환을 위한 재정 지원 및 직무 컨설팅 확대	-
	AI 실전형 산학협력 및 직무기반 훈련 시스템 구축	-
③ R&D	AI 격차 해소 위한 공공 오케스트레이션 플랫폼 공개	-
④ 기타	AI 직무 재교육 및 평생학습 프로그램 확대	수용성 (AI 리터러시, 역량 향상 교육)
	1인 기업·AI 기반 프리랜서 창업 지원	융합 및 혁신(창업촉진)과 연계성 있으나 노동시장 재편 대응 관점
	오케스트레이션형 교육 커리큘럼 도입	수용성, 시민사회 (AI 오케스트레이션 교육)

〈주제별(개인정보 보호/활용) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	고위험 분야 AI 결정권 제한 및 사용자 개입 의무화	책임, 관계 (인간 개입 및 책임 소재 명확화)
	데이터 제공/보호의 자기 결정권	개인정보보호/활용(기록 저장/활용)
② 제도, 행정	AI 에이전트 간 개인정보 교류 기준 및 인증제 도입	책임, 시민사회(AI 인증제) 관계(자격제)
	데이터 패스포트 제도 도입	-
	AI 에이전트 기억주기 설정권 및 인터페이스 구축	-
③ R&D	자기결정권 기반 기술 개발	개인정보보호/활용(자기 결정권)
	AI 블랙박스 기술 개발 (행위기록 저장 및 알림)	책임(사고추적, 기록, 분석 기술)

정책 분류	정책 제안	연관있는 주제(정책)
	개인정보 활용 위한 인터페이스 표준화 및 공공 데이터셋 구축	-
④ 기타	개인정보보호 리터러시 교육	
	통합 패스워드 관리 방안 마련	

〈주제별(사회적 수용성과 미래교육) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	AI 에이전트 공공접근 보장법 제정	-
	AI 책임소재 명확화 법제화	책임, 관계(법적 책임)
	AI 에이전트 인증제 및 표시제 도입	개인정보, 책임(인증제, 투명성)
	AI 교육법 및 교과 개발 (초·중등 리터러시)	책임, 시민사회(윤리 교육, 교육법)
② 제도, 행정	공공 AI 지식도우미 플랫폼 구축	관계, 시민사회(공공 접근성 향상)
	공공 MCP 허브 및 마켓플레이스 구축	융합및혁신(공공 테스트베드, MCP)
	AI 사회적 영향평가 제도화	시민사회, 책임(참여형 모니터링)
	AI 디지털교과서 및 교사 역량 강화	-
	AI 교사 인증제 및 '협력학습' 교과 추진	-
	AI 에이전트에 대한 편향 영향 평가 (BIA: Bias Impact Assessment)	
	한국적 문화 특성과 동양권의 문화 친화형 AI 에이전트 가이드라인	
	문화 다양성 반영 원칙 반영	
③ R&D	신뢰기반 지식 제공 AI 에이전트 기술개발	시민사회(정보 신뢰성 확보)
	인클루시브 AI 에이전트 기술 개발	-
	정보 검증 및 허위정보 차단 기술 개발	개인정보, 책임
	AI 교육 콘텐츠 및 효과분석 기술 개발	-
	핵심 역량 규명 및 직무별 시범사업	-

정책 분류	정책 제안	연관있는 주제(정책)
④ 기타	**AI 활용 시민 리터러시 교육 바우처 지급**	노동, 책임(AI 리터러시 교육 강조)
	중장년 대상, 지역 기반(격차해소) 생활형 AI 교육 강화	-
	AI-STEAM 연계 교육 및 직무 실습 커리큘럼	노동(직무 재교육)
	기술 숙련 기회 박탈 대응을 위한 비AI 공예시장 육성	-
	AI 중독 예방 및 거리두기 교육	-
	자율성·정체성 보호 교육 내용 강화	관계, 책임(자율성, 의존)

〈주제별(책임) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	**AI 에이전트 행위 결과의 책임 분담 체계 정비**	수용성, 관계(법적 책임 논의)
	개발 단계에서 설명 가능성 및 투명성 의무화	개인정보, 융합 및 혁신(XAI)
	고위험 분야 사전 허가제 및 활용 제한	개인정보, 수용성(고위험 AI 제한)
	자동화 수준 및 인간 개입 기준에 따른 활용 규제	관계, 책임, 수용성(인간 개입)
	AI 사고 피해 구제 위한 법적 장치 및 보험제도 도입	융합 및 혁신, 관계(보장 제도)
② 제도, 행정	AI 에이전트 개발자 책임 가이드라인 수립	윤리
	사고 대응 전담기구 설립	-
③ R&D	경찰 AI 에이전트 도입 (감시·개입)	-

〈주제별(윤리) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
② 제도, 행정	윤리 활용 가이드라인 제안	책임
	가치-윤리 충돌 해소를 위한 기술 기준 확보	-
③ R&D	사회적 윤리인식 변화 인지를 위한 모니터링 및 인공지능 학습 기술개발	-
	윤리적 판단 및 검출, 충돌 회피 기술개발	-

〈주제별(주권 및 안보) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	자율무기 규제 및 인간 통제 원칙 국제 협약	책임, 수용성(고위험 AI 규제)
	AI 무기 인권침해 방지 법제 및 감시 데이터 제한	개인정보, 책임 (감시 AI 규제, 로그 보존)
	AI 에이전트 등록제 및 로그 보관 의무화	책임, 개인정보 (행위 추적 블랙박스)
	반독점·상호운용성·데이터 이동성 의무화	융합 및 혁신 (개방형 생태계 조성)
	에이전트 로그 및 학습 데이터 국내 저장 의무화	개인정보(저장, 감사 시스템)
② 제도, 행정	국제 AI 거버넌스 규범 형성 주도	-
	공공 우선 도입 및 내수 확대 위한 도입 지원금	관계, 수용성 (공공 AI 플랫폼 구축 및 보급)
	오픈소스 생태계 및 기술 종속 방지 로드맵 수립	-
	공공 AI 클라우드·슈퍼컴퓨팅 인프라 구축	융합 및 혁신(인프라 지원)
	국산화 의무·산업별 보안 인증제 도입	책임, 융합 및 혁신(인증제)

정책 분류	정책 제안	연관있는 주제(정책)
③ R&D	AI 에이전트 국제 협약 체결 및 공동 R&D	-
	인간 중심 AI 감시시스템 및 오인식 최소화 기술 개발	-
	K-Agent 프로토콜 등 국산 표준 규격 개발	-
	멀티에이전트 오케스트레이터 독자 개발	융합 및 혁신(MCP 생태계조성)
④ 기타	AI 유니콘 육성 위한 투자 펀드 조성 및 규제 완화	융합 및 혁신 (창업 지원, 규제 샌드박스)
	AI 국부펀드 및 글로벌 AI 인재 유치	-

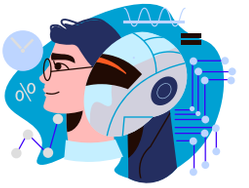
〈주제별(AI의 지속 가능성) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	신재생 에너지 연계 데이터센터 구축 및 법제화	-
② 제도, 행정	공공 조달 연계 인증 모델 우선 적용	수용성(공공 인증 AI 우선도입)
	AI 에너지 위원회 등 거버넌스 설립	-
	국제 AI 지속가능성 협력 및 공동연구 확대	-
③ R&D	저전력·경량 AI모델 개발 및 Green AI 인증제 도입	-
	데이터센터 전력수요 저감 기술개발	-
	AI 반도체 등 자원 순환형 R&D 추진	-
④ 기타	기후 AI 스타트업 육성 및 투자 확대	-
	AI 활용 탄소중립 제조공정 실증 지원	-

〈주제별(시민참여와 민주주의) 정책 제안〉

정책 분류	정책 제안	연관있는 주제(정책)
① 법, 규제	디지털 시민 토론 AI 윤리 가이드라인 제정	수용성, 책임(AI 윤리 가이드라인)
	공공 AI 에이전트 중립성 인증제 도입	책임, 개인정보, 수용성 (인증제)
	AI 정치 콘텐츠 생성·확산 제한 조항 신설	-
	정보 추천 알고리즘 다양성 기준 가이드라인	개인정보(정보 편향 방지)
	공인 뉴스 콘텐츠 우선 배치 제도화	-
② 제도, 행정	공공정책 참여용 AI 에이전트에 대한 제도 정비	-
	AI 콘텐츠 사전 등록제(선거기간)	-
	정보 다양성 공개 및 알고리즘 감사제	-
	전통언론-AI 간 공동 팩트체크 네트워크 구축 (감시 모델, 플랫폼 수행을 법적 규정)	책임(정보 검증)

3. 기술영향평가 정책 제언



주제별 정책 수요

AI 에이전트 각 이슈에 대응하기 위해 10대 주제에 대해 ① 법·규제 30개, ② 제도·행정 37개, ③ R&D 22개, ④ 기타 19개, 유사 중복 포함 등 108개 정책 수요가 도출되었다. 데이터 활용, 책임 소재 등 법적 쟁점 대응을 위한 법·규제가 제시되었고 제도적인 문제로는 사회 격변이 예상되는 노동, 사회적 수용성 분야의 직무 전환 재교육, AI 인증제 등 제도 마련이 제안되었다. AI 에이전트 활용 및 생태계 구축, 국제 표준 선점 등을 위한 산업 프로토콜, 공공 에이전트 등 R&D 제안이 있었고 수용성 향상, 윤리적 부작용 대응 중심의 리터러시 교육 등 기타 정책 제안도 있었다.

정책 우선순위 설정

정책 수요에 대한 기술영향평가 전문가 설문을 통해 우선순위에 따른 주요 정책을 제안하게 된다. 전체 정책에 대해 1차(상대), 2차(절대) 설문하여 전체 정책의 우선순위에 따라 주요 정책이 제안되는 것이다. 1차 설문은 108개 전체 정책에 대해 포괄적인 중요성을 기준으로 100점을 배분하는 상대평가인데 이를 통해 상위 23개 정책이 추려진다. 이때 유사·중복 정책 간 통합을 수행하며 정책 적용 시점, 중요도, 목적 등 각 주제별 특성을 고려하여 23개가 선정된다. 2차 설문은 1차 결과 추려진 상위 23개에 대해 2024년 안전·신뢰AI 평가 지표를 준용하여 절대평가를 수행한다. 여기서 상위 10개 주요 정책이 선별된다. 평가 지표는 중요성, 시급성, 효

과성, 실현 가능성 등으로 1~10점의 점수가 매겨진다.

〈(평가 지표) 중요성, 시급성, 효과성, 실현가능성(1~10점)〉

지표	내용
중요성	정책이 목표로 하는 문제의 규모와 심각성, 그리고 정책 시행으로 인한 사회적 영향력의 크기
시급성	정책적 개입의 시간적 제약성과 즉각적 대응의 필요성 정도
효과성	정책이 추구하는 방향과 결과가 사회적 가치와 규범에 부합하는 정도
실현 가능성	주어진 자원과 환경 하에서 정책을 실제로 수행할 수 있는 가능성의 정도

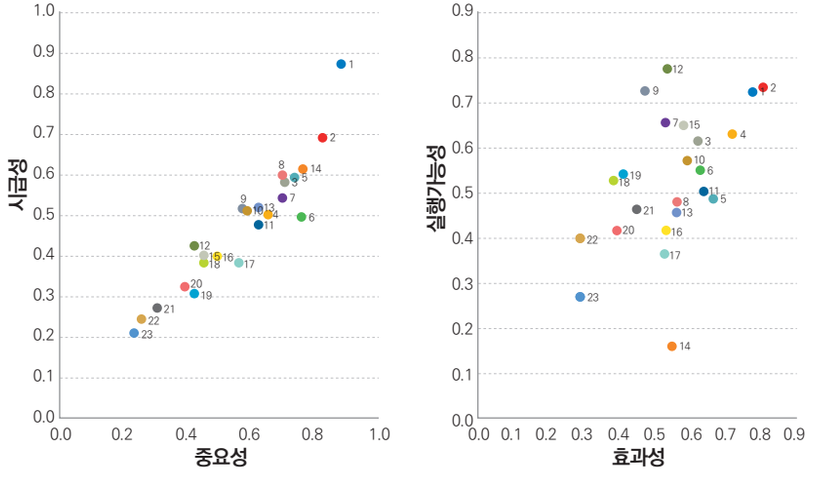
※ (표준화) 전문가 설문 점수는 지표별로 표준화* 후 분석

* (0-1 표준화) 최소값을 0, 최대값을 1로 변환하여 정규화

** (Z-score 표준화) 평균과 표준편차를 이용해 분포를 표준화

주요 정책 제안

〈전문가 정책 우선순위 평가〉



* 0-1 표준화에 따른 점수 분포로 각 지표별 최소 0 ~ 최대 1

책임 법제화, 교육 강화 등 AI 에이전트의 자율성과 행동에 대응하기 위한 정책적 대안이 우선순위로 확인되었다(붙임3).

〈정책 제안 순위 및 총점 현황〉

* 각 지표 원점수를 0-1 표준화 후 합산

순위	정책 제안	총점*
1	AI 에이전트 책임 법제화 및 책임 분담 체계 구축	3.25
2	AI 활용 및 협력 역량 강화 교육	3.05
3	데이터센터 전력수요 저감 기술개발	2.52
4	AI 사회적 영향 모니터링 확대	2.49
5	직무 전환을 위한 재정 지원 및 직무 컨설팅 확대	2.47
6	고위험 분야 AI 허가제 및 개입 의무화	2.43
7	산업 특화 AI에이전트 서비스 개발 및 실증	2.42
8	멀티에이전트 보안 프레임워크 구축	2.33
9	AI 에이전트 규제 샌드박스 운영	2.29
10	AI 에이전트 블랙박스 기술 개발 및 등록제	2.26

중요성, 시급성, 효과성, 실현 가능성 등을 종합적으로 고려하여 상위 10개의 정책을 주요 정책으로 분석하였다. 각 이슈에 대응하기 위해 제안된 정책들은 모두 필요성이 인정되지만 우선 추진이 필요한 주요 정책을 선정했다.

〈주요 정책 제안과 관련 부처〉

순번	주요 정책 제안	관련 부처
1	AI 에이전트 책임 법제화 및 책임 분담 체계 구축	행정안전부, 개인정보보호위원회, 과학기술정보통신부
2	AI 활용 및 협력 역량 강화 교육	교육부, 과학기술정보통신부, 고용노동부
3	데이터센터 전력수요 저감 기술개발	기후에너지환경부, 산업통상부, 과학기술정보통신부
4	AI 사회적 영향 모니터링 확대	과학기술정보통신부
5	직무 전환을 위한 재정 지원 및 직무 컨설팅 확대	고용노동부
6	고위험 분야 AI 허가제 및 개입 의무화	행정안전부, 과학기술정보통신부
7	산업 특화 AI에이전트 서비스 개발 및 실증	산업통상부, 중소벤처기업부
8	멀티에이전트 보안 프레임워크 구축	산업통상부, 개인정보보호위원회, 과학기술정보통신부
9	AI 에이전트 규제 샌드박스 운영	산업통상부, 과학기술정보통신부
10	AI 에이전트 블랙박스 기술 개발 및 등록제	행정안전부, 과학기술정보통신부

(1) AI 에이전트 책임 법제화 및 책임 분담 체계 구축

AI 에이전트 책임 법제화는, AI 에이전트의 활동에 따른 책임 소재를 명확히 하고 통제 불능 상황에 대한 통제, 결과에 대한 책임 분담 체계를 구축하는 것을 말한다.

AI 에이전트와의 관계

자율성이 부여된 AI 에이전트와 에이전트 간 상호 작용의 확대에 따른 통제 불능 우려가 예상된다. 잘못된 의사 결정으로 인한 법적 책임 소재가 불명확하며 AI 에이전트 간 경쟁적 상호 작용에서 사기, 권한 오남용의 위험도 존재한다. AI가 사용자의 의지에 반하거나(대리인의 딜레마) AI 군집 간 상호 작용에 의한 통제 불능 사태가 되는 것에 대비하여 법적·제도적 대응 체계를 구축할 필요가 있다.

사회적 수용성과 미래 교육

AI 에이전트의 결정 과정이 불투명하거나 '의도'를 알 수 없으면 이용자들이 시스템을 불신하게 되어 수용도가 떨어진다. 필요할 때는 'AI 에이전트 책임법'을 제정해야 하며 이 법 시행령에 AI 에이전트의 법적 지위와 책임 소재를 명확히 규정하는 내용을 포함시켜 불확실성을 해소해야 한다. AI 에이전트가 잘못된 의사 결정을 했을 때 사용자, 개발자, 플랫폼 중 명확한 책임 주체 및 배상 체계 등을 법률로써 명시해야 한다. 고위험 분야의 AI 에이전트에 대해서 EU 사례를 참고하여 강화된 안전성 기준과 감독 체계를 마련하여 적용할 필요가 있다.

책임

AI 에이전트의 사고와 피해 대응은 단일 제도로는 한계가 있으며, 시민보험–민간보험–법적 책임으로 이어지는 다층적 보장 체계가 필요하다. AI 관련 사고의 보험 통계(actuarial statistics) 자료의 신속한 구축을 위해 정부 차원의 AI 에이전트 사고 보고 체계를 구축해야 한다. 또 고위험 AI 서비스에 대한 인증·감독제 도입 및 민간보험 의무 가입 제도화가 필요하면 경미한 피해는 시민안전보험과 같은 보편적 사회보험으로 신속하게 처리하여 사회적 불안을 최소화해야 한다. 중범위 피해의 경우에는 민간보험을 활용하여 개발사·운영사·사용자가 책임을 분담하고, 보험시장의 신

산업으로 활성화할 필요가 있다.

대규모 피해나 인명 피해와 같이 중대한 사건에 대해서는 법적 책임을 강력하게 적용함으로써 최종 안전망을 확보하고 사고가 발생했을 때, EU AILD 사례처럼 AI에 관한 전문지식을 갖추지 못한 피해자가 아닌 개발자·운영자·사용자가 책임 설명을 입증해야 한다.

일정 수준의 안전성과 투명성을 확보한 경우 개발자, 미 RISE 사례와 같은 운영자에게 조건부 면책을 제공하는 등 인센티브 제도를 도입하는 것도 고려할 만하다. 장기적으로는 독립 법률 체계를 구축할 필요도 있다. AI 책임보험법은 일정 규모 이상의 사업자에게 배상책임보험 가입 의무화하는 것이고, AI 안전법 데이터 보존, 로그 기록, 사고 보고 의무, 안전 설계 기준 등을 규율하는 등의 법이다.

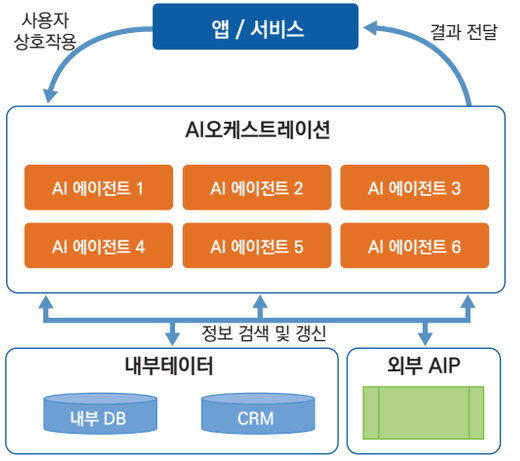
(2) AI 교육 강화

'AI 교육'이란, AI 리터러시 및 오케스트레이션, 격차 해소 등을 위한 교육 강화와 관련 지원 제도 확대 추진을 말한다.

노동 변화

다중 에이전트 협업으로 발전함에 따라 사용자가 여러 AI 에이전트의 작업을 지휘·조율하는 오케스트레이터로 전환될 것이다. 오픈 AI의 CEO 샘 알트만은 AI를 기반으로 1인 유니콘 기업이 등장이 가능할 것으로 예측했다. 오케스트레이션 교육·인증–창업·일자리 선순환 구축으로 알고리즘 기반의 창업이 활성화될 것도 예상된다.

〈AI 오케스트레이션 개념도(Sendbird, KB경영연구소 재구성)〉



사회적 수용성과 미래 교육

AI 활용 능력의 차이에 따라 개인 간 성과 격차가 기하급수적으로 커질 수 있으며, 기술이 기존 격차를 확대하고 고착화할 위험이 있다. 부유한 집이 모여 있는 학군의 학교들은 AI 도입 및 활용에 적극적인 반면 재정이 열악한 학교들은 준비 부족으로 출발선부터 격차가 벌어지는 양상이 나타나고 있다. 고령층, 장애인, 다문화 가정 등 취약 계층 대상 디지털 역량 강화 교육인 '디지털배움터' 사업을 전국민 대상 AI 교육으로 확대할 필요가 있다.

AI 바우처를 발급하여 국민 모두가 사각지대 없이 AI 서비스를 활용할 수 있도록 지원하고 AI 디지털 리터러시 교육과 연계할 필요가 있다. AI 바우처를 통해 AI 유료서비스를 이용하기 위해서 필수적으로 AI 리터러시 교육을 이수하고, 교육 이수 시 디지털 지갑에 보관 가능한 디지털 배지를 발급하는 제도를 시행할 만하다. 기존 법(지능정보화기본법 제23조(전문

인력의 양성), AI기본법 제21조(전문인력의 확보), 교육기본법 제23조(교육의 정보화)) 뿐만 아니라 교육기본법에 인공지능 교육 및 윤리 조항을 신설하여 AI 윤리 교육 및 리터러시 증진을 위한 관련 교육을 의무화해야 한다.

제23조의4(인공지능 교육 및 윤리) (신설)

① 국가와 지방자치단체는 모든 국민이 인공지능 기술을 올바르게 이해하고 활용 할 수 있도록 인공지능 리터러시 교육에 필요한 시책을 수립 · 실시하여야 한다.

② 국가와 지방자치단체는 인공지능 기술의 개발과 활용 과정에서 발생할 수 있는 윤리적 문제에 대응하기 위하여 인공지능 윤리 교육을 의무화하여야 한다.

③ 제1항과 제2항에 따른 교육에는 인공지능의 작동 원리, 사회적 영향, 편향성과 차별 방지, 개인정보 보호, 책임 있는 사용법 등이 포함되어야 한다.

(3) 데이터센터 전력 수요 저감 기술 개발

AI 확대로 인한 데이터센터 전력 수요가 증가할 것으로 예상되며 전력 사용 저감을 위한 최적화 R&D 추진이 필요하다.

AI의 지속가능성

상시 동작·지속 통신 구조의 AI 에이전트 특성으로 인해 연산·저장·네트워크 사용이 누적되어 전력·냉각·물·자원 수요 증가가 예상된다. 에너지 측면에서는 신재생 연계 데이터센터, 탄소 인식 스케줄링, 액체 냉각·폐열 회수 등 고효율 인프라 전환이 가속화될 것으로 보인다. AI는 학습 시 방대한 컴퓨팅 자원을 필요로 하며 필연적으로 많은 양의 전력을 소모하게 되는데 예를 들면 오픈 AI GPT-3 모델 학습에 1,287MWh 전력을 소비하고 550t의 CO_2 배출량을 기록(Wired 2023)했다. 이 전기 소모량은 미국의 약 120가구 1년치 사용량과 같다.

엣지 컴퓨팅 등을 활용한 저전력·경량 AI모델 개발을 장려하고, AI모델의 효율성을 측정하는 라벨링 제도 도입이 필요하다. AI 데이터센터의 전력 소비량을 줄일 수 있는 관리 최적화 시스템 및 학습 시 전력 소모를 줄이는 알고리즘 R&D 사업도 추진해야 한다.

〈AI 관련 글로벌 전략 수요 전망〉

(KISTEP, AI로 인한 전략수요의 폭발적 증가와 대응방안, 2025)

구분	기준	현재 수요	예상 수요	성장률
국제에너지기구 (IEA)	글로벌 데이터센터	460TWh (2022)	1,000~2,000TWh (2030)	117~335%
세미애널리시스 (SemiAnalysis)	글로벌 AI 데이터센터	49GW (2023)	74~132GW(2028)	196%
보스턴컨설팅그룹 (BCG)	글로벌 AI 데이터센터	60GW (2023)	127GW(2028)	212%
맥킨지(McKinsey)	글로벌 AI 데이터센터	55GW (2023)	171~219GW (2030)	311~398%
골드만삭스 (Goldman Sachs)	글로벌 AI 데이터센터 (암호화폐 제외)	400TWh (2023)	1,040TWh(2030)	260%
전략국제문제연구소 (CSIS)	미국 AI 데이터센터	4GW (2024)	84GW(2030)	2,100%

(4) AI 사회 영향 평가 제도화

AI 에이전트 확대에 따른 우리 사회의 영향을 지속적으로 관찰 및 대응하기 위한 사회 영향 평가가 수행되어야 한다.

AI 에이전트와의 관계

AI 에이전트에 의한 과도한 최적화로 경험 영역이 축소되는 인지적 갇힘 현상(echo chamber)이 발생할 우려가 있다. AI 에이전트 도입에 따른 인간 관계 및 정체성의 변화에 대한 문화적·사회적 이해 확산도 필요하다. 인간-에이전트 관계 변화에 따른 사회적 영향 평가와 문화 연구 지원이 확대되어야 한다. 감정노동·돌봄노동·교육 등 영역에서 AI 활용이 인간관계에 미치는 변화를 분석하고 인간의 행위성 및 정체성과 사회 관계 재구성에 대한 연구를 활성화할 필요가 있다.

사회적 수용성 및 미래 교육

AI 에이전트는 정책·산업·교육 등 공공성이 높은 영역에 깊이 관여하므로 신뢰성·투명성·책임성 확보가 사회적 수용성의 전제가 된다. 정부, 기업, 전문가, 시민단체 등이 모여 AI 에이전트 도입에 따른 사회적 영향을 모니터링하고 갈등을 조정할 필요가 있다. 과기정통부는 2022년부터 AI 서비스 확산에 따른 사회적 영향을 살펴보고 사전 대응하기 위해 'AI 서비스 사회적 영향 평가'를 통해 조사·분석하고 있다. AI 에이전트 도입 과정에서 발생하는 이해관계 충돌을 선제적으로 해소하고 발생하는 긍정적/부정적 영향을 조사하고 분석하는 것이다.

(5) 직무 전환을 위한 재정 지원 및 직무 컨설팅 확대

이는 AI로 인한 구조적 실업이 예상되는 직종 및 근로자를 대상으로 직무 전환을 지원하는 제도를 마련하는 것이다.

노동 변화

AI 에이전트는 일부 직종에서 노동력 대체를 가져올 것으로 예상되며 근로자의 스킬 노후화가 가속화될 것으로 보인다. 골드만삭스는 AI가 정규직 일자리 3억 개 대체한다고 보고[10]했고, 한국개발연구원은 2030년에는 국내 일자리 열 개 가운데 아홉 개에서 AI가 70% 이상의 업무를 대체할 수 있다고 보고했다.[11]

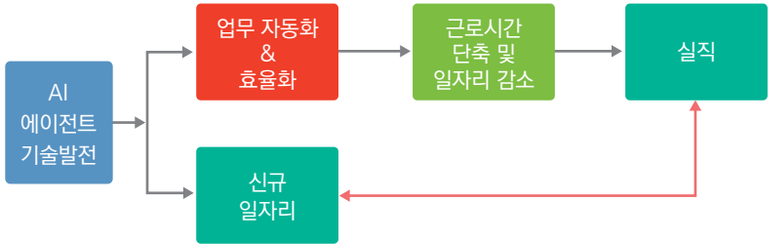
AI 에이전트가 업무 보조를 넘어 노동력을 대체하면서, 일부 직종에서는 대규모 구조적 실업이 발생할 가능성이 증가하고 있다. 대체할 가능성이 큰 직종에 교육 수준과 전문성이 높은 직종이 다수 포함되어, 기존 일자리 정책의 변화가 요구된다. 마이크로소프트가 발표한 AI로 인해 위험에 처한 직종에는 역사가, 번역가, 데이터 과학자, 수학자, 에디터 등 대학 이상의 고등교육을 요구하던 전문직이 다수 포함되어 있다.[12]

일자리를 위협받는 직종 종사자들의 저항으로 사회적 갈등이 유발될 가능성이 있다. 구조적 실업이 예상되는 직종의 근로자에게 재교육, 직무 전환 등을 통해 새로운 업무나 직종에로의 적응 유도가 필요하다. 단기적으로 다양한 AI 에이전트를 능숙하게 다룰 수 있을 정도로 심화된 근로자에게는 재교육을 제공하는 것이 필요하다. 중기적으로 전문직이 큰 영향을 받으므로 대학(원)에서 더 다양하고 전문성 높은 AI 직무 재교육에 대한 재정을 제공할 필요가 있다. (가칭) 'AI 직무 재교육 사업' 지원 등이다. 싱가폴 스킬스퓨처(SkillsFuture) 프로그램은 국공립대학, 기술 교육 기관 등에서 기술 재교육을 받을 때 드는 비용을 최대 90%까지 지원하며, 이들 다양한 기관을 통해 광범위한 AI 관련 교육 프로그램을 제공하고 정부는 이를 인증하는 시스템이다.

근로자 재교육이나 직무 전환으로 인한 기업의 부담을 경감시킬 수 있

도록, 중소 사업체 위주로 세제 혜택 등을 통한 지원을 제공할 필요가 있다. 싱가폴 SkillsFuture 프로그램의 경우 개인뿐 아니라 중소 사업체에도 근로자 교육 프로그램 이수 및 직무 전환 때문에 발생하는 비용을 지원한다. 일부 직종에서 AI 에이전트로 인한 구조적 실업이나 산업 대체가 불가피하므로, 이에 대한 선제적 정책 개입이 필요하다.

〈AI 에이전트로 인한 일자리 양극화〉



(6) 고위험 분야 AI 허가제 및 개입 의무화

이는 고위험 분야에서 AI 활용 시 사용자의 위험 인지나 사고 예방, 책임 판단 등을 위한 법이나 규제를 마련하고자 함이다.

개인 정보 보호와 활용

소규모의 데이터를 활용하는 경우 식별 가능성이 크지 않지만, 점차 개인형 서비스가 창출되며 많은 데이터를 조합하게 되면 개인 특정이 가능해진다. 금융, 의료 등 개인 정보를 활용한 고위험 분야에서의 AI 활용은 위험 요소가 크므로 사용자의 권리와 책임 규정이 필요하다. 고위험 분야에서는 AI 에이전트의 결정권 범위를 제한하고 필요에 따라 사용자 승인을 의무화할 필요가 있다. 사용 여부 판단 등 데이터 소유자의 자기결정권 보장이 요구되며, 자기결정권에 기반한 사용자의 개입과 판단을 수행할 수 있다.

책임

의료·군사·금융 등 고위험 활용에서 오류·오작동이 생겼을 때 막대한 피해나 사회적 비용이 예상되므로 사전 규율과 단계적 보장 제도를 연계할 필요가 있다. 위험도 평가 기준을 수립하고 인증제·허가제 도입 검토하며 고위험 서비스 제공자는 민간보험 가입을 의무화하여 중범위 사고를 대비해야 한다. 대규모 사고나 인명 피해가 발생했을 때는 강력한 민·형사의 법적 책임을 병행하여 최종 안전망을 마련할 필요가 있다.

(7) 산업 특화 AI 에이전트 서비스 개발 및 실증

이는 심화되는 글로벌 경쟁에 대응하기 위해 산업 분야에 특화된 맞춤형 에이전트 서비스 기술 개발과 실증을 말한다.

융합 및 혁신

구글, 오픈AI 등 글로벌 기업이 AI 에이전트 시장을 선도하고 있으나, 스타트업의 오픈 소스를 토대로 한 경쟁도 치열한 상황이다. 지난 2년간 AI 에이전트 관련 스타트업 투자 자본은 20억$(약 2.78조 원)을 넘어섰다.[13] 국내 유망 스타트업·중소기업 등에 다양한 산업적 기회를 제공하여 글로벌 AI 에이전트 활용 시장을 선점할 필요가 있다. 기존 생성형 AI 시장이 AI 에이전트 시대로 급변함에 따라 기술 개발이나 상용화 시장 선점을 위한 R&D 및 실증 사업 확대도 필요하다.

특화 파운데이션 모델, 설명 가능한 인공지능, eX플레인블(eXplainable) 인 AIXAI, 신뢰성 및 안정성 확보 등을 위한 기술 개발 추진이 필요하며

다양한 산업 분야에 특화된 맞춤형 에이전트 서비스 개발·실증 혹은 실증 사업 지원도 추진되어야 한다. AI 에이전트 활용 B2B/B2C 서비스 실증 지원이 필요한데 AI 에이전트의 산업 적용을 위한 실증은 PoC 단계로 글로벌 시장 경쟁이 치열하다.

(8) 멀티에이전트 보안 프레임워크 구축

이는 멀티 에이전트 활용에 대비하여 거버넌스 체계를 수립하고 관련 보안 프레임워크를 구축하는 것을 말한다.

융합 및 혁신

다수의 에이전트가 상호 작용하는 복잡한 네트워크 구조로 인해 단일 취약점이 전체 시스템에 연쇄적 피해를 일으킬 가능성이 증가했다. 특정 프로토콜에 대한 업계 의존도가 높아질 경우, 장애, 보안 결함 등이 동시 영향을 미치는 시스템적 위험도 증가한다. 여러 에이전트의 복잡한 상호작용으로 알고리즘 편향이나 오류가 발생할 때 원인 규명이나 책임 소재 파악이 곤란할 수 있다.

또한 에이전트 간 자동화된 협업 과정에서 인간의 감독이나 개입 시점에 대한 명확한 정의가 곤란하여 예상치 못한 상황에 대한 대응력이 떨어질 수 있다. 그래서 멀티에이전트 보안 프레임워크 구축을 위한 거버넌스 체계를 수립할 필요가 있다. 분산형 시스템 보안 인증, 에이전트 간 신뢰 체인 구축, 연쇄 장애 방지를 위한 격리 메커니즘도 의무화해야 한다.

(9) AI 에이전트 규제 샌드박스 운영

규제 샌드박스 운영은 AI 에이전트 확산을 위해 실시된다.

융합 및 혁신

AI 에이전트는 기술 발전뿐 아니라 하드웨어, 클라우드 인프라 등 다양한 기술적, 경제적 성장 기회를 창출할 전망이다. 그래서 국제 표준 개발 참여나 MCP 등 글로벌 표준(de facto) 대응을 위한 환경을 조성할 필요가 있다. MCP, A2A 등 빅테크 중심 에이전트 기술과 국내 상호 운용할 수 있는 생태계가 구축되어야 하고 산업 현장 내 적용 유도를 위한 규제 샌드박스 적용과 법 정비 등 기업의 컴플라이언스 대응을 위한 지원도 필요하다.

금융 마이데이터 등을 활용한 실시간 AI 에이전트 서비스를 개발하기 위해 개인 정보 처리 등에 대한 현행 제도를 개선할 필요가 있다. 규제 샌드박스 종료 후 지속성 등 한계를 극복하기 위해 개인정보보호법, 저작권법, 전자금융거래법 등의 관련법 정비도 필요하다.

〈AI 에이전트 관련 주요 관련 법·규범〉

관련 법/제도	주요 내용
개인정보보호법	데이터 활용 규제 완화 및 명확화
정보통신망법	AI 자율 의사결정 시 책임 주체 불명확
산업안전보건법	인간, 에이전트 협업 환경에서 안전 규제 미비
상법	AI 대리행위 관련 명확성 부족
전자금융거래법	AI가 금융 의사결정 및 거래 수행가능성 근거 부족
국제규범(EU AI Act 등)	글로벌 규제 정합성 확보 필요(고위험 AI 분류 평가체계 등)

(10) AI 에이전트 등록제 및 블랙박스 기술 개발

이는 AI 안전성과 윤리 확보를 위한 AI 에이전트 의무 등록과 행위 감시 제도를 마련하는 것을 말한다.

개인 정보 보호와 활용

AI 에이전트의 고도화로 인해 사용자의 민감한 정보가 장기적으로 수집·공유될 가능성이 증가하고 있다. 이에 개인의 프라이버시 침해와 정보 남용을 방지하기 위한 법적·제도적 장치를 마련하는 것이 시급하다. AI 에이전트의 행위를 실시간으로 감시하기 위한 AI 블랙박스 기술을 개발할 필요가 있고 AI 에이전트의 모든 행동과 질문, 응답 기록을 블랙박스처럼 저장하고 침해 소지 있는 행위를 모니터링하며 반드시 사용자에게 알리고 동의를 구하도록 해야 한다.

주권 및 안보

AI 안전성과 윤리의 국제 표준을 선도하고, AI 에이전트의 책임 있는 활용을 위한 법적·제도적 체계를 구축해야 한다. 모든 AI 에이전트 의무 등록제 도입하고 행위 로그 의무 보관을 통한 추적 가능성을 확보하여 AI 에이전트 행위 체계를 구축해야 한다. AI 에이전트의 자율성을 수준별로 차등 규제하는 체계를 도입하고 위험성이 높은 AI에 대해서는 인간 개입을 의무화할 필요가 있다.

집필에 참여해주신 분들

성명	소속	직책
이재신	중앙대학교	교수
강동오	한국전자통신연구원	책임연구원
김태원	한국지능정보사회진흥원	수석
박종열	서울과학기술대	교수
안성원	소프트웨어정책연구소	실장
임화섭	한국과학기술연구원	단장
조명수	정보통신산업진흥원	팀장
장하영	써로마인드	대표
허상우	네이버	연구원
박희제	경희대학교	교수
이상욱	한양대학교	교수
이승철	서울대학교	교수
이충환	동아에스앤씨	이사

연구에 참여해주신 분들

구분	성명	소속	직책
과학기술 정보통신부	조시훈	과학기술정보분석과	과장
	장종덕	과학기술정보분석과	서기관
	이은숙	과학기술정보분석과	주무관
한국과학 기술기획평가원	신동평	기술예측센터	센터장
	오건웅	기술예측센터	연구위원
외부연구진	박성연	크리베이트파트너스 (시민포럼 기획/진행)	
	양정은	크리베이트파트너스 (시민포럼 기획/진행)	

시민포럼

성명	분야
김O	사회
김OO	문화
송OO	환경
심OO	외교
윤OO	기술
전OO	경제
조OO	예술
홍OO	안보

주석

1 기술 범위는 학계 및 산업계에서 통용되는 개념이 아니며 본 기술영향평가 추진을 위해 설정된 기술 범위임

2 TechCrunch, “Common Sense: 72% of US teens have used AI companions”, 2025-07-21.

3 IBM, “X-Force Threat Intelligence Index 2025”, 2025.

4 Melissa Kay Diliberti et al. (2024)., Using Artificial Intelligence Tools in K-2 Classrooms.

5 김혜미, “인공지능(AI)의 법적 책임 귀속에 관한 연구”, 경희대학교 법무대학원 석사학위논문, 2025.

6 “Parents of teenager who took his own life sue OpenAI.” BBC News . August 27, 2025. https://www.bbc.com/news/articles/cgerwp7rdlvo

7 “Airline held liable for its chatbot giving passenger bad advice- what this means for travellers.” BBC News. February 23, 2024. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know.

8 “I let ChatGPT’s new ‘agent’ manage my life. It spent $31 on a dozen eggs.” Washington Post. February 7, 2025. https://www.washingtonpost.com/technology/2025/02/07/openai-operator-ai-agent-chatgpt/

9 “Tesla found partly liable in 2019 autopilot death.” Wired . August 1, 2025, https://www.wired.com/story/tesla-liable-2019-autopilot-crash-death/.

10 BBC News, 2023.3.30.

11 동아일보, 2024.7.16.

12 Fortune, 2025.7.31.

13 딜로이트 2025 첨단기술·미디어·통신산업 전망

발행일 2026년 7월 000000000000000일

발행처 과학기술정보통신부, 한국과학기술기획평가원(KISTEP)

ISBN 979-11-