3. 인공지능 분야의 윤리 협회

이전 섹션에서 자세히 설명한 것처럼 인공지능(AI)의 개발, 사용, 그리고 효과에 대한 수많은 고려 사항들이 존재한다. 한 사회 내에서 인공지능이 시민의 기본 인권에 미칠 수 있는 잠재적 영향부터 수집된 데이터의 보안과 활용에 이르기까지, 동일한 개발자 집단에 의해 뜻하지 않게 내장된 편견과 차별부터 주어진 AI의 선택과 활용의 결과에 대한 대중의 인식과 이해의 부족에 이르기까지의 범위는 잘못된 결정과 추가적인 손해로 이어진다.

AI는 과거의 ICT 및 컴퓨팅 혁명에 기반하여 구축되었으며, 이전과 마찬가지로 여러 비슷한 윤리적 문제에 직면하게 될 것이다. 선(善)을 위해 사용될 수 있는 기술이지만, 오용될 가능성도 있다. 우리는 인간과 기계 사이의 경계를 흐리면서 AI를 과도하게 의인화하고 인간화할 수 있다. 지속되는 AI의 발전은 새로운 '정보 격차'를 야기할 것이고, 다른 집단보다도 일부 사회경제학자와 지리학자 집단에 더 많은 혜택을 줄 것이다. 더욱이, AI는 생물권과 환경에 검증되지 않은 영향을 미칠 것이다.

3.1 국제 윤리 협회

공식적인 규정은 아직 부족하지만, 이러한 윤리적 난제를 탐구하기 위해 국제적으로 여러 독자적 협회들이 출범되었다. 이 섹션에서 알아보게 될 협회들은 표 3.1에 요약되어 있으며, 인공지능에 대한 대중의 이해와 완화를 목적으로 관련 결함과 문제를 고려하여 연구를 진행하게 될 것이다.

표 1 : 윤리협회와 제기된 위해

협회	장소	주요 이슈	간행물	자금 출처
인공지능윤리연구소	독일	인간 중심 공학인 철학, 윤리학, 사회학, 정치학 등의 학문을 포함하여 AI의 빠른 발전에 대한 문화적, 사회적 고착에 집중함		페이스북의 초기(2019년) 자금 지원(5년간 750만 달러)
AI&머신러닝 윤리 연구소	영국	이 연구소는 책임 있는 기계 학습을 위한 8가지 원칙을 바탕으로 개인에서 전 국가에 걸쳐 AI를 개발할 수 있는 권한을 부여하는 것을 목표로 함. 이러한 원칙은 인간 제어 유지, AI 영향에 대한 적절한 교정조치, 편향 평가,		미상

		명시성, 투명성, 재현성, 근로자에 대한 AI 자동화의 영향 완화, 정확성, 비용, 개인 정보 보호, 신뢰 및 보안에 관심을 가짐.		
인공지능윤리교육연구소	영국	젊은 세대에 대한 잠재적 위협과 새로운 AI 기술의 급속한 성장에 대한 교육, 그리고 AI가 주도하는 EdTech의 윤리적 발전을 보장함		미상
미래생명연구소	미국	자율 무기 군비 경쟁, AI의 인간 통제, 진보된 '일반/강력' 또는 초지능형 AI의 잠재적 위험 등 안전성과 실존하는 위험에 초점을 맞춰 AI의 개발이 인류에게 유익함을 보장함	아실로마 AI 원칙	개인 투자자 : 일론 머스크(스페이스X·테슬라), 자안 탈린(스카이프), 맷 웨이(금융 트레이더), 니산 스티엔논(소프트웨어 엔지니어), 샘 해리스, 조지 고둘라(기술 기업가), 제이콥 트레페텐(하버드)
전산기계협회	미국	연구, 개발 및 구현 측면에서 컴퓨터와 네트워크의 투명성, 사용성, 보안, 접근성, 책임성 및 디지털 포괄성	알고리즘 투명성과 책임성에 대한 설명(2017. 01), 컴퓨팅 및 네트워크 보안(2017. 05), 사물 인터넷(2017. 06), 접근성, 사용성, 디지털 포괄성(2017. 09), 법 집행을	미상

			위한 정보 인프라에 대한 의무 액세스(2018.04)	
일본AI학회	일본	AI R&D가 인간사회에 이롭게 남고, 개발과 연구가 윤리적이고 도덕적으로 이뤄지도록 보장하는 것	윤리 가이드라인	미상
AI4AII	미국	사회적 선과 인류의 이익을 위해 소수집단을 AI에 노출시키는 다양성과 AI의 포용력		구글
미래사회협회	미국	정책 연구, 자문 및 집단 지능, 거버넌스, 법률 및 교육에 걸쳐 광범위하게 사회에 이로운 인공지능의 영향과 거버넌스	2017년 10월 발간된 'AI 거버넌스 원칙 초안' (이후 2019년 2월 7일 홈페이지에 게재)	미상
AI현재연구소	미국	AI의 사회적 영향, 특히 권리와 자유, 노동과 자동화, 편향과 포함, 안전 및 중요 인프라와 같은 분야		루미네이트, 맥아더 재단, 마이크로소프트 리서치, 구글, 포드 재단, 딥마인드 윤리&소사이어티, AI 이니셔티브 윤리&거버넌스 등 다양한 기관
전기전자공학연구소	미국	AI 및 지능형 시스템을 인간 중심적으로 유지하고 인류의 가치와 원칙을 지원하기 위한 사회 및 정책 지침. 설계 및 개발 전반에 걸쳐 모든 이해 관계자가 인권, 웰빙, 책임, 투명성 및 오남용	윤리적 정렬 디자인(초판, 2019.03)	

		인식에 대한 윤리적 고려사항의 우선 순위를 지정할 수 있도록 교육, 교육 및 권한을 갖도록 하는 데 초점을 맞춤.		
AI파트너십	미국	AI 기술에 대한 모범 사례: 안전, 공정성, 책임성, 투명성, 노동 및 경제, 사람과 시스템 간의 협업, 사회적 사회적 영향, 사회적 이익		파트너십은 세계 6대 기술 회사를 대표하는 AI 연구원 그룹에 의해 형성되었다.Apple, Amazon, DeepMind Google, Facebook, IBM, Microsoft.
책임로봇재단	네덜란드	책임감 있는 로봇 공학(설계, 개발, 사용, 규제 및 구현 측면). 기술 혁신에 수반되는 문제 및 이러한 문제가 안전, 보안, 개인 정보 보호 및 웰빙과 같은 사회적 가치에 미치는 영향을 사전에 파악하는 것		미상
AI4People	벨기에	AI가 미치는 사회적 영향, 그리고 '올바른 AI 사회'를 구축하기 위한 설립 원칙, 정책, 관행	올바른 AI사회를 위한 윤리 프레임워크	아토미움 : 유럽 과학, 미디어 및 민주주의 연구소. 일부 자금은 공학 및 물리 과학 연구 위원회로부터 프로젝트의 과학 위원장으로부터 제공
AI협회윤리&거버넌스	미국	공정성, 인적 자율성 및 정의 등의 사회적		하버드 버크먼 클라인 센터와

		가치를 입증하는 방식으로 자동화 및 머신러닝 기술이 연구, 개발 및 배치되도록 함		MIT 미디어 랩. 마이애미 재단, 나이트 재단, 루미네이트, 레드 호프만, 윌리엄 앤 플로라 휴렛 재단의 지원을 받음
사이도트:책임 있는 AI 생태계 지원	핀란드	기업, 정부 및 조직이 책임 있는 AI 생태계를 개발하고 배치하여 투명하고 책임감 있고 신뢰할 수 있는 AI 서비스를 제공할 수 있도록 지원. 조직이 인간 중심의 AI를 개발할 수 있도록 지원하고, AI 생태계의 신뢰와 책임 수준을 높이는 데 초점을 맞춤. 플랫폼에서 '정보 시스템의 신뢰도를 검증할 수 있는 소프트웨어 및 알고리즘 시스템을 제공(Saidot, 2019).		
euRobotics	유럽	로봇 공학 분야에서의 유럽의 재능과 진보 유지 및 확장 - AI 산업화 및 경제적 영향		유럽집행위원회
데이터 윤리&혁신 센터	영국	규제 환경의 격차 파악 및 연결, AI의 데이터 사용, 사회에 대한 AI의 이점 극대화		영국 정부
SIGAI, 전산기계협회	미국	컴퓨팅 전반에 걸친 AI 원리 및 기술의 성장과 적용 촉진 및 지원, 다양한 포럼을 통한 AI 교육 및 간행물 홍보		전산기계협회

기타 주요한 국제 발전 : 현재와 과거				
몬트리올 선언	캐나다	사회 각 분야에 걸쳐 400명의 참가자들이 모여 장단기적으로 윤리적·도덕적 과제를 파악하는 AI의 사회 책임 발전. 주요 가치: 웰빙, 자율성, 정의, 프라이버시, 지식, 민주주의, 책임감.		몬트리올 대학교, 퀘벡 보건 연구 기금, 몬트리올 컨벤션
국제노동조합네트워크(UNI)	스위스	직장 내 AI, 로봇공학, 데이터 및 머신러닝 적용에 있어서의 작업자의 업무 중단과 투명성. 근로자의 이익 보호, 인적 통제와 건전한 힘의 균형을 유지.	AI 윤리 10대 원칙	미상
유럽로봇연구네트워크(EURON)	유럽(스웨덴 기반)	연구 협동, 교육 및 훈련, 출판 및 회의, 산업 연결 및 국제 로봇 연결	로봇 윤리 로드맵	유럽집행위원회(2000-2004)
유럽로봇플랫폼(EUROP)	유럽	유럽 로봇공학과 AI 커뮤니티를 하나로 묶음. 산업 중심으로 경쟁력 및 혁신에 초점		유럽집행위원회

3.2 국제 윤리 협회에 의해 제기되는 윤리적 문제점

앞서 나열된 모든 협회들은 AI가 윤리적인 방식으로 연구, 개발, 설계, 배치, 감시 및 사용되어야 한다는 데 동의한다. 그러나 각자 다른 분야를 우선시한다. 이 섹션에서는 협회들이 해결하고자 하는 문제에 초점을 맞춰 각 협회를 분석 및 분류하며, 위험성으로부터 보호하기 위해 제안된 접근법과 해결책에 대해 대략적으로 설명한다.

협회들의 주요 이슈는 크게 아래와 같은 범주로 나뉠 수 있다.

1. 인권과 웰빙
AI가 인간성와 인간의 건강한 삶을 위한 가장 큰 관심사가 될 수 있는가?

2. 정서적 피해

AI가 인간의 정서적 경험의 무결성을 저하시킬 것인가, 아니면 정서적 또는 정신적 피해를 촉진시킬 것인가?

3. 의무와 책임
누가 AI에 책임이 있고, 누가 AI의 행동에 책임을 질 것인가?

4. 보안, 개인 정보, 접근성 및 투명성
데이터와 개인화 서비스에 있어 접근성과 투명성, 사생활과 보안의 균형을 어떻게 유지할 것인가?

5. 안전과 신뢰
만약 AI가 대중들에게 신뢰를 얻지 못한다고 생각하거나, AI 혹은 다른 사람들의 안전을 위협하는 방식으로 행동한다면 어떻게 할 것인가?

6. 사회적 피해와 사회적 정의
AI가 포용력이 있고 편견과 차별이 없으며 공공 도덕과 윤리에 부합한다는 것을 어떻게 보장할 것인가?

7. 재정적 피해
경제적 기회와 고용에 부정적인 영향을 미치고, 인간 노동자의 일자리를 빼앗거나, 이러한 일자리의 기회와 질을 떨어뜨리는 AI를 어떻게 통제할 것인가?

8. 적법성과 정의
우리는 어떻게 AI와 그것이 수집하는 데이터가 정의롭고, 공정하고, 합법적이며, 적절한 통제와 규제의 적용을 받는 방식으로 사용되고, 처리되고, 관리되도록 할 것인가? 그러한 규제는 어떻게 보일까? AI가 '인격'을 부여받아야 하는가?

9. AI의 통제 및 윤리적 사용(또는 오용)
AI가 비윤리적으로 사용될 수 있는 방법은 무엇이며, 이로부터 인간을 어떻게 보호할 수 있는가? AI가 발전하고 '학습'하는 동안에도 AI가 완전히 인간의 통제 하에 있도록 어떻게 보장할 수 있는가?

10. 환경적 피해와 지속 가능성
AI의 개발 및 사용과 관련된 잠재적 환경 피해로부터 어떻게 보호하고 있는가? 지속 가능한 방법으로 어떻게 생산할 수 있을까?

11. 공공연한 사용
대중이 AI의 이용과 상호작용에 대해 인지하고, 교육받고, 알 수 있도록 하려면 어떻게 해야 할 것인가?

12. 현존하는 위험

어떻게 AI 무장 경쟁을 막고, 잠재적 피해를 선제적으로 완화하고 규제하며, 첨단 머신 러닝이 진보적이지만 관리하기 쉽도록 보장할 수 있는가?

전반적으로 협회들은 모두 인간의 이익을 가장 높은 수준에서 확립하고, 인간 사회와 환경 모두에 대한 편익을 우선시하며(이 두 가지 목표가 상충되지 않음) AI와 관련된 위험과 부정적인 영향을 완화하는 윤리적 프레임워크와 시스템을 식별하고 형성하는 것을 목표로 한다. 또한 AI가 책임감 있고 투명하다는 것을 확인하는 데 중점을 두고 있다.

IEEE의 '윤리적으로 조정된 디자인: 자율적이고 지능적인 시스템으로 인간의 행복을 우선시하는 비전'(v1; 2019)은 AI로부터 제기될 수 있는 윤리적 문제와 이를 완화하기 위해 제안된 다양한 방법에 대해 현재까지 발표된 가장 실질적인 문서 중 하나이다.

그림 2: 자율 및 지능형 시스템의 윤리 및 가치 기반 설계, 개발 및 구현에 대한 일반 원칙 (2019년 3월 IEEE의 윤리 정렬 설계 제1판에 의해 정의됨)

핵심 영향을 나타내는 영역은 지속 가능한 개발, 디지털 아이덴티티에 대한 개인 데이터 권리 및 기관, 책임에 대한 법적 프레임워크, 교육 및 인식 정책을 구성한다. 이러한 개념은 윤리적 조정 디자인 개념 프레임워크에서 보편적인 인간의 가치, 정치적 자기 결정과 데이터 기관, 그리고 기술적 의존성이라는 세 가지 축에 속한다.

3.2.1 세부적인 피해

이번 세션에서는 AI의 위해성이 협회에 의해 어떻게 개념화되어 있는지와 남아 있는 몇 가지 문제에 대해 탐구한다.

인권과 복지

모든 시책은 AI가 인간의 존엄성, 보안, 사생활, 표현의 자유, 개인정보 보호, 개인정보 보호, 평등, 연대, 정의 등 기본적이고 근본적인 인권을 침해해서는 안 된다는 견해를 고수하고 있다(유럽의회, 의회, 위원회, 2012).

우리는 어떻게 AI가 그러한 기본적인 인권을 유지하고 인간의 복지를 우선시하도록 할 것인가? AI가 아동, 장애자, 노인 등 사회 취약 계층에 불균형적인 영향을 미치거나 사회 전체의 삶의 질을 떨어뜨리지는 않는가?

인권보호를 보장하기 위해 IEEE는 AI 사용을 감독하는 새로운 거버넌스 프레임워크, 표준 및 규제 기관, 기존의 법적 의무를 정보에 입각한 정책으로 전환하고 문화적 규범과 법적 틀을 허용하며 AI에게 인간과 같은 권리와 특권을 주는 것을 허가하지 않고 항상 인간이 통제를 유지할 것을 권고한다. IEEE는 '인간의 삶의 만족과 삶의 조건에 대한 만족, 긍정적 영향과 부정적 영향 사이의 적절한 균형'으로 정의되는 인간 복지를 보호하기 위해 설계 단계 전체에 걸쳐 인간의 복지를 우선시하고, AI의 사회적 성공을 명확하게 측정하기 위해 가장 널리 수용되는 가용 지표를 사용할 것을 제안한다.

책임감과 투명성 사이에는 교차점이 존재한다. 권리 침해를 확인하고 추적하며 적당한 보상과 개혁을 제공할 수 있는 적절한 방법이 항상 존재해야만 한다. 개인정보 또한 중요한 이슈

중 하나이다. AI가 모든 종류의 개인정보를 수집하기 때문에, 사용자는 자신의 기본권이 합법적으로 유지될 수 있도록 데이터에 접근하고 제어할 수 있어야 한다(IEEE, 2019).

책임로봇재단(Foundation for Responsible Robotics)에 따르면 AI는 '책임로봇'이라는 목표를 달성하기 위해 윤리적으로 개발돼야 한다. 이는 안전, 보안, 개인정보보호, 웰빙 등 사회적 가치를 지키기 위한 선제적 혁신에 의존한다. 재단은 정책 입안자와 협력하고, 행사를 조직하고, 행사를 주최하고, 정책 입안자와 대중을 교육하기 위한 협의 문서를 발행한다. 또한 산업과 소비자 사이의 격차를 해소하고, 투명성을 높이기 위해 민간 협력을 창출한다. AI가 출시되어 활용되기 전에 이뤄지는 연구개발 단계부터 윤리적인 의사결정권, 소비자 교육 확대, 책임 있는 법·정책 수립이 요구된다.

미래생명연구소(Future of Life Institute)에서는 인간의 존엄성, 권리, 자유, 문화적 다양성에 대한 이상과 양립할 수 있는 방식으로 AI를 설계하고 운영해야 하는 필요성 등 AI의 발전에 따른 여러 원칙과 윤리, 가치관을 규정하고 있다. 일본AI윤리지침협회(Japanese Society for AI Ethical Guidelines)에서는 연구자와 사회 전체의 윤리와 양심, 역량에 따라 AI가 인류에게 유익한 방식으로 실현되어야 한다는 것을 중시하고 있다. ‘AI는 사회의 평화, 안전, 복지, 공익에 기여해야 하며 인권을 보호해야 한다’는 것이 이 협회의 의견이다.

미래법사회협회(The Future Society's Law and Society Initiative)에서는 인간은 번영할 권리, 존엄성, 자유가 평등하며, 인권을 누릴 권리가 있다고 강조한다. 그렇다면 우리는 사람들에게 영향을 미칠 수 있는 결정을 기계에게 어느 정도까지 위임해야 할까? 예를 들어, 법률 분야에서 AI가 인간보다 더 효율적이고, 공정하며, 균등하고, 비용 절감 효과가 있을 수 있는 ‘판결’을 내릴 수 있을까?, 그렇다고 하더라도 이것이 AI를 효율적으로 사용할 수 있는 적절한 방법일까? 몬트리올 선언은 AI의 출시에 영향을 받는 분야에서의 인권증진을 위해 국제적인 윤리 체계를 짜서 이를 어느 정도 명확히 하는 것을 목표로 한다. '현재 선언의 원칙은 인간이 감각과 사고, 감정을 부여받은 사회적 존재로서 성장하려고 하고, 감정적, 도덕적, 지적 능력을 자유롭게 발휘하여 잠재력을 실현하기 위해 노력한다'는 공통된 믿음에 달려 있다. AI는 인간의 행복을 저해해서는 안되고 개선 및 성장시킬 수 있도록 적극 권장하며 지원해야 한다.

국제노동조합네트워크(UNI Global Union)와 같이, 개인의 노동권을 보호하는기 위해 보다 구체적인 관점에서 AI에 접근하는 단체도 있다. 이 연합에서는 현재 사람들이 하고 있는 작업의 절반 이상은 자동화된 방법으로 더 빠르고, 더 효율적으로 수행할 수 있다고 설명한다. 이는 AI가 인간 고용의 영역에 가져올 명백한 피해를 나타낸다. 또한 연합에서는 AI가 인간과 지구에 도움이 될 수 있다는 것을 보장해야 하며, 기본적 인권, 인간의 존엄, 성실성, 자유, 프라이버시, 문화와 젠더의 다양성을 보호하고 높여야 한다고도 주장하고 있다.

정서적 위해

인간이란 무엇인가? AI는 아직 검증받지 못한 방식으로 인간의 감정적 경험에 상호작용하고 영향을 미칠 것이다. 인간은 긍정적으로나 부정적으로나 감정적으로 영향을 받기 쉽다. 또한 감정과 욕망이 어떻게 행동에 영향을 미치는지는 지능의 핵심 영역이다. 이러한 영향은 문화에 따라 다르며, 문화적 감수성이나 교류 방식을 고려하면 영향력 있는 AI가 사람들의 사회 그 자체의 관점에 영향을 줄 수 있다. IEEE는 이러한 위험을 줄이기 위해 예를 들면, AI의 규범과 가치를 상대에 따라 다르게 적용하고 갱신하는 기능이나, 문화에 대한 민감성 등 다양한

방법을 권장하고 있다.

AI가 정서적으로 위해를 가할 방법은 다양하다. 잘못된 친밀감, 과잉 애착, 신체의 객관화·상품화, 사회적·성적 격리 등이 있다. 이는 책임로봇재단(the Foundation for Responsible Robotics), AI파트너십(Partnership on AI), AI현재연구소(the AI Now)과 같은 감성 컴퓨팅과 관련된 협회들과 몬트리올 선언(the Montréal Declaration), 유럽로봇연구네트워크(European Robotics Research Network, EURON) 로드맵(휴머노이드의 위험에 관한 섹션을 진행하는 로드맵)과 같은 다양한 윤리 협회들에 의해 다뤄지고 있다.

이러한 위해는 성산업에서 AI와의 친밀한 관계 발전을 고려할 때 두드러지게 나타난다. IEEE에 따르면 친밀한 시스템(Intimate system)은 성차별, 인종적 불평등 또는 신체적으로 부정적인 고정관념에 기여해서는 안되고, 긍정적이고 치료를 위한 방향으로 사용되어야 하며, 사용자의 동의가 없는 성적 또는 심리적 조작을 피해야 한다. 또한 사용자가 인간 동료로부터 고립되도록 설계해서는 안 된다. 인간관계의 유동성과 질투심에 미치는 영향에 대해 투명한 방법으로 설계되어야 하고, 일탈행위나 범죄행위를 조장하거나 소아성애나 강간 같은 불법적인 성적 관행을 정상화해서는 안 되며, 상업적으로 거래되어서는 안 된다(법률적인 의미에서).

정서적 AI가 사용자를 속이고 협박할 수 있는 가능성 또한 존재한다. 연구자들은 AI가 사용자를 정서적으로 조작하고 영향을 줄 때, 행동을 미묘하게 수정하는 행위를 '너징(nudging)'으로 정의했다. 이것은 약물 의존성, 건강한 식습관과 같은 면에서는 유용할 수 있지만, 인간의 건강을 악화시키는 행동을 유발할 수도 있다. 시스템 분석을 위해서 AI를 개발하기 전에 감성 설계의 윤리를 검토해야 한다. 사용자는 너징을 인식하고 구분하는 방법에 대해 교육을 받아야 하고, 자율 너징 시스템을 위한 사전동의 시스템을 갖춰야 한다. 어린이와 같이 정보에 입각한 동의를 할 수 없는 취약 계층은 추가적인 보호를 받아야 한다. 전반적으로 책임자들은 이기적이거나 해로운 용도에 쓰일 수 있는 AI의 너징 설계 경로가 윤리적인 경로인지 아닌지에 대한 문제를 논의해야 한다(IEEE, 2019).

IEEE(2019)에 의해 제기된 바와 같이, 너징은 공공 행동에 영향을 미치기 위해 정부와 다른 기업에 의해 사용될 수 있다. 예를 들어 로봇이 자선 행위나 기부를 장려하기 위해 너징을 사용하는 것이 윤리적으로 적절한가? IEEE는 오용의 가능성으로 인해 그러한 행동의 수혜자에 대해 완전히 투명성을 추구해야 한다고 말한다.

기술 중독과 사회적 또는 성별 편견으로 인한 정서적 해악과 관련된 문제들도 존재한다.

책임과 의무

대부분의 협회는 AI를 감사할 수 있어야 한다고 규정하고 있다. AI의 설계자, 제조자, 소유자 및 운영자가 기술적 조치와 이로 인해 발생할 수 있는 잠재적 위해에 대해 책임을 지도록 하기 위해서이다. IEEE에 따르면, 이는 법원이 개발 및 배치 단계에서 책임 소지를 명확히 함으로써 관계자들이 그들의 의무와 권리를 이해할 수 있도록 함으로써 달성될 수 있다. 또한 설계자와 개발자는 다양한 사용자 그룹들 사이에서 존재하는 문화적 규범의 다양성을 고려할 수 있고, AI 중심의 기술이 최신 기술인 점을 감안하여 현재 존재하지 않는 규범을 만들고, 특정 AI에 대해 법적 책임이 있는 사람을 상시 추적할 수 있도록 등록 및 기록보관 시스템을 구축하기 위해 다중 이해 당사자(multi-stakeholder) 생태계를 구축할 수 있다.

미래생명연구소(Future of Life Institute)에서는 장단기적으로 윤리성을 갖추기 위해 AI가 따라야 할 23가지 지도원칙의 목록인 '아실로마 원칙(Asilomar Principles)'을 통해 책임에

관한 문제를 다룬다.

첨단 AI 시스템의 설계자와 구축자는 '사용, 오용, 행동의 도덕적 이해 당사자로, 그러한 함의를 구체화할 책임과 기회를 가지고 있다'(FLI, 2017). AI가 실수를 해야 한다면 그 이유가 무엇인지도 설명할 수 있어야 한다. AI파트너십(Parnership on AI)도 편향성 측면에서의 책임의 중요성을 강조하고 있다. 우리는 가정(assumption)과 편향(bias)이 데이터 내부와 데이터로 구축된 시스템 내에 존재한다는 사실에 민감해야 하며, 이를 복제하지 않도록 노력해야 한다. 즉, 공정하고 편향이 없는 AI를 구축하기 위해 적극적으로 책임을 져야 한다.

다른 모든 협회들도 설계자와 AI 엔지니어, 그리고 규정, 법률 및 사회 전반에 걸쳐 책임의 중요성에 대해 강조하고 있다.

성(Sex)과 로봇

2017년 7월, 책임로봇재단(Foundation for Responsible Robotics)은 '로봇과 함께하는 우리의 성적 미래'(Foundation for Responsible Robotics, 2019)에 관한 보고서를 발표하였다. 이것은 기술과의 친밀한 연관성을 둘러싼 다양한 이슈와 의견들에 대한 객관적인 요약을 제공하는 것을 목표로 했다. 많은 나라들이 성적 만족을 위해 로봇을 개발하고 있는데, 대부분 인체를 포르노적으로 표현한 경향이 있고, 여성형 로봇이다. 이렇게 사람처럼 의인화된 결과물은 성적 만족감이 친밀감, 우정 및 대화의 요소와 결합될 때 로봇이 생물과 무생물 사이의 무언가로 인식되도록 할 수 있다. 로봇은 또한 성별이나 신체 고정관념에 대한 사회적 인식에 영향을 미칠 수 있으며, 사람간의 연결과 친밀감을 없애고 사회적 고립을 더욱 심화시킬 수 있다. 그러나 로봇이 성범죄를 줄이는데 도움을 주고, 치료요법에 이용하여 강간이나 성적 학대의 피해자의 재활에 사용하는 등 인간에게 정서적으로 성적 이익을 줄 수 있다는 가능성도 존재한다.

접근성과 투명성 vs. 보안과 개인정보 보호

AI에 대한 주요 관심사는 투명성, 명시성, 보안성, 재현성 및 해석성이다. 즉, 시스템이 특정한 결정을 내린 이유와 방법, 또는 로봇이 그 방식으로 행동하는 이유와 방법을 발견할 수 있을지에 관한 것이다. 이는 특히 신체 손상에 직접적인 영향을 미칠 수 있는 안전 중요 시스템(예: 무인 자동차 또는 의료 진단 시스템)의 경우 매우 중요하다. 투명성이 존재하지 않는다면 사용자는 자신이 사용하고 있는 시스템 및 관련 결과를 이해하거나 관련자에게 책임을 묻기가 어려워질 것이다.

이를 해결하기 위해 IEEE는 측정과 시험이 가능한 투명도의 수준을 상세히 기술하는 새로운 표준을 개발하여 시스템의 준수 여부를 객관적으로 평가할 수 있도록 했다. 이는 이해 관계자에 따라 다른 형태를 취할 수 있다. 로봇 사용자는 '작동이유(why-did-you-do-that)' 버튼을 요구할 수 있으며, 인증 기관이나 사고 조사자는 고장 원인을 투명하게 제공하는 '윤리 블랙박스(ethical black box)' 형태로 관련 알고리즘에 접근할 수 있을 것이다(IEEEE, 2019).

AI는 자동 의사결정 방식을 지속적으로 학습하고 발전시키기 위해 데이터를 필요로 한다. 이러한 데이터는 개인 데이터이며 특정 개인의 물리적, 디지털 또는 가상 ID(개인 식별 가능 정보, PII)를 식별하는 데 사용될 수 있다. IEEE에서는 '결과적으로 모든 디지털 트랜잭션(명시되거나 관찰된)을 통해 인간은 자신의 물리적 자아에 고유한 디지털 그림자를 생성한다(2017)'라고 설명했다. 인간은 특정 정보를 비공개로 유지할 권리를 어느 정도까지 실현할 수 있는가? 또는 이러한 데이터가 어떻게 사용되고 있는지 알 수 있는가? 사람들에게는 자신의 고유한 정체성을 통제하고 배양하며 데이터 사용에 관한 윤리적 영향을 관리할 수 있는 적절한 도구가 부족할 수 있다. 교육받지 않았다면 많은 AI 사용자들은 스스로가 남기고 있는 디지털 발자취와 그들이 세상에 내보내고 있는 정보에 대해 여전히 모르고 있을 것이다. 사용자가 데이터를 제어하고 상호 작용 및 액세스할 수 있도록 시스템을 배치하고 디지털 페르소나를 통해 대신 접근할 수 있도록 해야 한다.

PII는 개인의 자산으로서 인정받았으며(유럽에서 규정(EU) 2016/679에 의해), 개인의 자율성, 존엄성 및 동의권을 보호하기 위해 데이터가 수집되고 사용되는 시점에 명시적 동의를 요청해야 한다. IEEE는 개인이 스스로 개인정보를 제어하고 머신 러닝 데이터 교환의 잠재적 윤리적 영향을 예측하고 완화하기 위해서 개인화된 '프라이버시 AI 또는 알고리즘 요원'의 도입 가능성에 대해 언급했다.

미래생명연구소(Future of Life Institute)의 아실로마 원칙은 IEEE의 다양한 측면에 걸친 투명성과 프라이버시의 중요성에 대해 동의한다. 아실로마 원칙은 오류 투명성(AI가 오류를 낸다면, 그 이유를 밝혀낼 수 있어야 한다), 사법 투명성(사법적 의사결정에 관여하는 모든 AI는 인간에게 만족스러운 설명을 제공해야 한다), 개인 프라이버시(AI가 수집하고 생성하는 데이터에 사람이 액세스, 관리 및 제어할 수 있는 권한이 있어야 한다), 자유와 프라이버시(AI가 사람들의 실제 또는 인식된 자유를 부당하게 축소해서는 안 된다)로 구성된다. 사이도트(Saidot)는 조금 더 넓은 접근 방식을 취하며, 협력·진보·혁신을 위해 사람·기관·스마트시스템이 쉽게 연결되어 협력하는, 투명하고 책임감 있고 신뢰할 수 있는 AI의 중요성을 강하게 강조한다.

관련 협회들은 모두 AI의 투명성과 책임성을 중요한 이슈로 파악하고 있다. 이러한 밸런스는 법적, 사법적 공정성, 근로자 보상 및 권리, 데이터 및 시스템의 보안, 공공 신뢰, 사회적 해악과 같은 우려를 뒷받침한다.

자율성과 행위자 vs. 피동자

현재 AI에 대한 접근은 부인할 수 없이 인간중심적이다. 이것은 도덕적 행위자(moral agents)와 도덕적 피동자(moral patients) 사이의, 인공적이고 자연적이며, 자기 조직적인 것과 그렇지 않은 것 사이의 구별에 관해 가능한 문제들을 제기한다. AI는 생명체가 자율적이라고 간주되는 것과 같은 방식으로 자율화될 수 없지만(IEEE, 2019), AI의 관점에서 자율성을 어떻게 정의해야 하는가? 기계의 자율성은 기계가 규칙에 따라 어떻게 작동하고 작동하는지를 결정하지만, 감정과 도덕성을 AI에 이식하려는 시도는 '행위자과 피동자의 구별을 흐리고 기계를 의인화할 수 있다는 기대감을 조장할 수 있다'고 IEEE는 설명한다. 특히, 구현된 AI가 인간과 점점 더 비슷하게 보이기 시작하고 있기 때문이다. 인간과 시스템/기계 자율성 사이에 구별을 설정하는 것은 자유 의지, 존재 및 사전 결정의 문제를 포함한다. 인공지능과 시스템 측면에서 '자율성'이 무엇을 의미할 수 있는지를 명확히 하기 위한 추가 논의가 필요한 것은 분명하다.

안전과 신뢰

AI가 인간의 의사결정을 보완하거나 대체하기 위해 사용되는 상황에서는 안전하고, 신뢰성이 높고, 신뢰할 수 있어야 하며, 청렴해야 한다는 공감대가 형성돼 있다.

IEEE는 연구자들 사이에서 ‘안전 사고방식’을 구축하고, ‘의도적이지 않고 갑작스러운 행동을 식별 및 예방하고’, ‘안전한’ 시스템을 개발할 것을 제안한다. 프로젝트 및 진행 상황을 평가하기 위한 자원 및 수단으로 기관에 검토 위원회를 설치하고, 공동체가 공유하도록 장려하며, 안전 관련 개발, 연구 및 도구에 대한 소식을 널리 알리는 것이다. 미래생명연구소(Future of Life Institute)의 아실로마 원칙(Asilomar Principles)에서는 AI의 개발과 배치에 관여하는 모든 사람이 임무 주도적이어야 하며, AI는 ‘한 국가나 조직이 아닌 널리 공유된 윤리적 이상을 서비스하고, 모든 인류의 이익을 위해 개발되어야 한다’는 규범을 채택해야 한다고 명시하고 있다(2017). 이러한 접근 방식을 통해 AI에 대한 대중의 신뢰를 쌓을 수 있을 것이고, 성공적인 사회 통합의 핵심이 되어 줄 것이다.

일본AI학회(the Japanese Society for AI)에서는 AI는 항상 진실되게 행동해야 하며, AI와 사회가 서로 배우며 진지하게 의사소통을 해야 한다고 제안한다. 이 협회에서는‘지속적이고 효과적인 커뮤니케이션’은 상호간에 이해를 높이고 '인류 전체의 평화와 행복에 공헌할 수 있다'(JSAI, 2017)고 저술했다. AI파트너십(Partnership on AI)은 AI의 신뢰성을 보장하고 AI 과학자와 기술자 간 협력, 신뢰, 개방의 문화를 구축하기 위해 노력하고 있다. AI윤리&머신러닝 연구소(Institute for Ethical AI&Machine Learning)는 대화의 중요성을 강조한다. 또한 8개의 핵심이념에서 신뢰와 사생활 문제를 연계시켜 AI 기술자가 관련된 프로세스 및 데이터에 대해 관계자와 소통하고 신뢰를 쌓아 사회 전체에 이해를 확산시킬 것을 의무화하고 있다.

사회적 해악과 사회적 정의: 포괄성, 편견, 차별

AI 개발에는 다양한 관점이 필요하다. 이러한 관점들이 사회적 관점과 일치하고 사회적 규범, 가치관, 윤리관, 선호를 따라야 하며, 데이터나 시스템에 편견과 가정이 구축되어서는 안 되며, AI가 문화의 다양성을 존중하는 공공의 가치, 목표, 행동에 따라야 한다는 규범을 확립하고 있는 협회들이 존재한다. 또한 이 협회들은 모든 사람이 혜택을 받을 수 있도록 AI가 공공의 이익을 위해 기능해야 한다고 주장하고 있다. 바꿔 말하면, AI의 개발자나 사용자에게는 AI에 적절한 가치관을 심어놓아서 AI가 사회의 어떤 부분에 대해서도 현재 혹은 장래에 걸쳐 해를 끼치거나 악화시키지 않도록 해야 할 사회적 책임이 있다는 것이다.

IEEE는 먼저 AI가 배치될 특정 커뮤니티의 사회적, 도덕적 표준과 AI가 제공할 특정 작업이나 서비스의 표준을 식별하고, 서비스 표준은 정적이지 않다는 점과 AI가 문화에 따라 역동적이고 투명하게 변화해야 한다는 점을 고려하여 '표준 업데이트'에 관한 아이디어를 염두에 두고 AI를 설계하고, 이를 식별할 것을 제안한다. 그리고 사람들이 통상적으로 갈등을 해결하는 방법을 인식하고 그것을 AI에 비슷하고 투명한 방법으로 실현하는 시스템을 장비하는 방법 또한 제안하고 있다. 이는 특정 사회집단에 불리한 잠재적 편견을 고려하여 주의를 기울이면서 협력하여 다양한 연구에 걸쳐 이루어져야 한다.

AI4All 및 AI현재연구소(the AI Now)와 같은 여러 단체들은 모든 단계에서 공정하고 다양성을 추구하며 차별 없는 AI를 명시적으로 지지하며, 소수 집단에 대한 지원에 초점을 맞추고 있다. 현재 AI 관련 학위 프로그램은 예비 개발자와 설계자에게 적절한 윤리 지식을 갖추고 있지 않고(IEEE, 2017), 기업 환경과 비즈니스 관행 또한 윤리에 대한 도움이 되어주지 못하고 있어 가치 기반 혁신을 주도하고 지원할 수 있는 선임 윤리학자가 부족한 상태이다.

세계적인 규모로 보면 선진국과 개발도상국 간의 불평등 격차가 상당하다. AI가 인도주의적 측면에서 상당한 유용성을 가질 수 있지만, 이러한 나라 간 격차를 심화시키거나 빈곤, 문맹, 성별, 인종 불평등을 악화시키고 고용과 노동을 불균형적으로 방해해서는 안 된다. IEEE는 불평등 격차 완화를 위해 조치와 투자하는 것, 기업의 사회적 책임(CSR)을 개발과 마케팅에 통합하는 것을 제안한다. 또한 투명한 전력 구조 개발, 로봇공학 및 AI에 관한 연구 촉진 및 공유, AI를 미국의 지속 가능한 개발 목표에 부합하도록 만들 것을 제안한다. AI 기술은 글로벌 표준화 및 오픈 소스 소프트웨어를 통해 전 세계적으로 동등하게 이용할 수 있도록 해야 하며, 효과적인 AI 교육 및 훈련에 대한 학제간 논의가 이루어져야 한다(IEEE, 2019).

일본AI학회(the Japanese Society for AI)가 발표한 일련의 윤리지침은 무엇보다도 인류에 대한 기여와 사회적 책임을 강조한다. AI는 공공의 이익을 위해 행동해야 하고, 문화적 다양성을 존중하며 항상 공정하고 일관적인 방법으로 사용되어야 한다.

책임로봇재단(Foundation for Responsible Robotics)은 책임감 있는 AI를 추진함에 있어 '다양성에 대한 헌신'을 포함한다. AI파트너십(Parnership on AI)은 데이터 안에 숨겨진 편견과 가정들의 존재를 무시하는 '심각한 사각지대'에 대해 주의를 주고 있다. 사이도트(Saidot)는 우리의 사회적 가치가 '점점 알고리즘에 의해 조정되고 있다'고 하지만, AI는 여전히 인간 중심적으로 작동되고 있다(2019). 미래생명연구소(Future of Life Institute)는 문화적 다양성과 인권이라는 인간적 가치가 배어 있는 AI의 필요성을 강조하고 있으며, AI윤리&머신러닝 연구소(Institute for Ethical AI&Machine Learning)는 AI 개발과 생산에서의 편향을 모니터링하기 위한 '편향성 평가'를 제시한다. 인간의 편견과 가정이 가지고 있는 위험은 AI의 지속적인 개발과 함께 자주 확인되는 위험이다.

재정적인 위험 : 경제적 기회와 고용

AI는 경제에 피해를 입히고 많은 사람들의 일자리를 없애거나 업무를 중단시킬 수 있다. 그리고 많은 종류의 업무가 자동화되고 관련 업계가 변화하면서 사라지게 됨에 따라 노동자의 권리와 교체 전략에 영향을 미칠 것이다.

또한, 단순히 일자리를 잃거나 얻는 숫자에 초점을 맞추기보다는, 자동화의 효과를 완화하고 고용의 복잡성을 고려하기 위해서 전통적인 고용 구조를 변화시켜야 한다. 기술 변화가 너무 빠르게 일어나고 있기 때문에 기존 노동자들은 재교육 없이는 업무를 따라가지 못하고 있다. IEEE(2019)는 근로자는 일에 적응하기 훈련해야 하고, 재교육을 받을 수 없는 사람들을 위한 폴백 전략이 적용되어야 하며, 향후 고용에 대한 접근성을 높이기 위해 훈련 프로그램을 고등학교 이하의 수준으로 구성해야 한다고 말한다. 국제노동조합네트워크(UNI Global Union)는 디자이너, 제조업체, 개발자, 연구자, 노동조합, 변호사, CSO, 건물주, 고용주 등이 한자리에 모이는 글로벌 및 지역 차원의 다중 이해 당사자(multi-stakeholder) 윤리 AI 거버넌스 기구를 만들 것을 요청하고 있다. AI는 경제적, 기술적, 사회적 디지털 격차를 해소하고 근본적인 자유와 권리를 지지하여 공정한 전환을 보장하는 정책을 시행함으로써 사람들에게 널리 평등한 혜택을 주고 힘을 실어 주어야 한다.

AI현재연구소(The AI Now Institute)는 다양한 분야의 고용과 근로조건의 성격을 바꾸는 자동화와 초기 단계 통합 등 AI가 노동과 업무에 미칠 영향을 더 잘 이해하기 위해 다양한 이해관계자 그룹과 함께 연구한다. 미래사회협회(Future Society)는 AI가 법조계에 어떤 영향을 미칠지 구체적으로 의문을 제기한다. 'AI 시스템이 법률 업무의 특정 측면에서 인간 변호사보다 우수하다면, 법률의 실천에 있어 윤리적이고 전문적인 함의는 무엇인가?' (미래학회, 2019)

AI는 직장 내에서 근로자의 임금에 보다 더 많은 영향을 미칠 것이며, 다양한 긍정적인 기회를 제공할 수도 있을 것이다. IEEE(2019)에 의해 규정된 바와 같이, 작업장의 편향에 대한 잠재적 해결책을 제공할 수 있으며(위에서 언급한 바와 같이 이를 염두에 두고 개발된 경우에), 제품 개발의 결함을 찾아냄으로써 설계 단계에서 사전 예방적으로 개선할 수 있다.

RRI는 혁신 프로세스와 시장성 있는 제품에 대한 윤리적 허용 가능성, 지속 가능성, 사회적 만족도에 관해 우리 사회에 과학적, 기술적 진보를 가져오기 위해서 사회적 행위자들과 혁신자들이 서로 협력하는 투명하고 상호작용적인 프로세스이다(Von Schomberg, 2013).

책임있는 연구와 혁신(RRI) RRI는 프로젝트 초기부터 윤리적 우려를 해결할 수 있는 도구를 제공하기 위해 고전 윤리로부터 파생되는, 특히 EU에서 성장하고 있는 분야이다. RRI가 프로젝트의 설계 단계에 통합되면, 그 설계가 윤리적 지지 측면에서 목적에 적합하고 강경하게 될 가능성이 높아진다. 많은 연구 기금과 조직들은 RRI를 그들의 연구 및 혁신 과정에 포함시키고 있다(IEEE, 2019).

합법성과 정의

여러 협회들은 AI가 합법적이고 공정하며, 정의롭고, 적절하고, 선제적인 거버넌스 및 규제의 대상이 되어야 할 필요성에 대해 연구하고 있다. AI를 둘러싸고 있는 복잡한 윤리적 문제들은 직간접적으로 분리된 법적 도전으로 전환된다. AI는 제품으로서 어떻게 라벨링되어야 할까? 동물? 사람? 아니면 그 외의 새로운 무언가로?

IEEE는 AI에 어떠한 수준의 '인격'도 부여해서는 안 되며, AI의 개발, 설계, 유통은 모든 해당 국제법과 국내법을 완전히 준수해야 하지만 관련 법률을 정의하고 이행하는 데에는 많은 노력이 필요하다고 결론짓는다. 법적 문제는 법적 지위, 정부 사용(투명성, 개인의 권리), 피해에 대한 법적 책임, 투명성, 책임 및 검증 가능성 등 몇 가지 범주로 나뉜다. IEEE는 AI가 재산법의 적용을 받아야 한다고 꾸준히 주장한다. 이해관계자는 AI에게 맡겨서는 안되는 결정들의 유형을 구별해야 하며, 규칙과 기준을 통해 그러한 결정을 인간이 효과적으로 통제할 수 있도록 보장해야 한다. 또한 기존 법률을 면밀히 검토하고 AI에 법적 자율성을 부여할 수 있는 메커니즘을 검토해야 한다. 그리고 개발자와 운영자는 AI가 작동할 수 있는 모든 사법권 영역에서의 해당 법률을 준수해야 한다. 또한 IEEE는 AI가 점점 더 정교해지고 있기 때문에 정부 차원에서 AI의 법적 지위를 재평가하고, 규제 기관, 사회 및 산업 행위자 및 기타 이해관계자들과 긴밀히 협력하여 AI 시스템의 개발이 아닌 인류의 이익을 우선시할 것을 권고하고 있다.

AI의 통제와 윤리적 사용 및 오용

정교하고 복잡한 AI가 개발될수록 오용 가능성은 높아진다. 개인정보는 악의적인 의도로 이익을 위해 사용될 수 있고, 시스템이 해킹당할 위험성이 있으며, 기술이 공격을 위해 사용될 수 있다. 그래서 AI가 무엇인지 알고 사용해야 하고, 대중적 인식도 중요하다. 우리가 새롭게 AI 시대에 접어들면서, 지금까지 없었던 새로운 시스템과 기술이 등장했기 때문에 일반 시민들도 이러한 AI의 사용이나 오용으로 발생할 수 있는 위험에 대한 최신 정보를 알고 있는 것이 좋다.

IEEE는 대중들에게 윤리 및 보안 문제에 대해 교육하는 새로운 방법을 제안한다. 예를 들어 개인정보를 수집하는 스마트 기기에 대한 '데이터 프라이버시' 경고가 있다. 또한 확장 가능하고 효과적인 방법으로 이 교육을 제공하고, 경찰관이 학교에서 안전교육을 하는 것과 마찬가지로 정부 관계자와 국회의원 및 집행 기관을 교육하여 시민들과 협력적으로 일하고 AI 보안에 대한 공포감이나 혼란을 피할 수 있게 할 수 있도록 해야 한다고 제안하고 있다(IEEE, 2019).

행동과 데이터의 조작에 관한 문제도 존재한다. 인간은 AI에 대한 통제권을 유지하고 반란을 거부할 수 있어야 한다. 대부분의 협회들은 이것이 AI가 발전함에 따라 직면하게 될 잠재적 문제로 규정하고 있다. 그리고 AI가 예측 가능하고 신뢰할 수 있는 방식으로 행동해야 하며 적절한 해결 수단을 갖추고 검증 및 테스트 대상이 되어야 한다고 강조한다. 또한 AI는 인류를 위해 일해야 하고, 사람을 착취해서는 안 되며, 인간 전문가들에 의해 정기적으로 검토되어야 한다고 주장한다.

인격과 AI

AI가 '인격'을 가질 자격이 있는지에 대한 문제는 의무감, 자율성, 그리고 책임감을 둘러싼 논쟁으로 연결된다. AI의 행동과 결과에 책임이 있는 것은 AI 그 자체인가, 아니면 그것을 만든 사람인가?

이 개념은 인간적인 의미에서 로봇을 사람으로 간주하는 것을 허용한다기보다는, 기업과 같은 법적 수준으로 간주할 뿐이다. 기업의 법적 인격이 현재 법의 영향으로부터 인간을 보호할 수 있다는 점에 주목할 필요가 있다. 다만 국제노동조합네트워크(UNI Global Union)는 법적 책임은 로봇 자체가 아니라 개발자에게 있다고 주장하며, 로봇에 책임을 전가하는 것을 금지해야 한다고 요청한다.

환경 오염과 지속 가능성

AI의 생산, 관리, 구현은 지속 가능하고 환경에 피해를 입히지 않아야 한다. 이는 '웰빙'의 개념과도 관련이 있다. 웰빙의 주요 인식 측면은 대기, 생물 다양성, 기후 변화, 토양 및 수질 등 환경과 관련되어 있다(IEEE, 2019). IEEE(EAD, 2019)는 AI가 지구의 자연 시스템에 해를 끼치거나 환경오염을 악화시키지 않아야 하며, 지속 가능한 관리, 보존 또는 지구 자연 시스템의 복원을 실현하는 데 기여해야 한다고 명시한다. 국제노동조합네트워크(UNI Global Union)은 AI가 사람과 행성을 최우선으로 해야 한다고 주장하며, 지구의 생물 다양성과 생태계를 보호하고 향상시키기 위해 노력하고 있다. 책임로봇재단(Foundation for Responsible Robotics)은 농업과 농업의 역할부터 기후 변화 모니터링, 멸종 위기 종의 보호에 이르기까지 AI의 다양한 잠재적 용도를 파악하고 있다. 재단은 리스크를 완화하고 지속적인 혁신과 개발을 지원하기 위해 AI와 로봇을 통제하는 것에는 책임감 있고 정보에 입각한 정책이 필요하다고 말한다.

정보에 입각한 사용 : 공교육과 인식

국민 구성원은 시민 참여, 소통, 국민과의 대화를 통해 AI의 이용, 오남용, 잠재적 해악에 대해 교육을 받아야 한다. 동의의 문제(그리고 개인이 얼마나 합리적이고 그것에 대해 알고 있는가)가 이것의 핵심이다. 예를 들어, IEEE는 동의가 윤리적인 것보다 명확하지 않은 몇 가지 사례를 제기한다. 만약 개인정보가 불쾌하거나 알지 못하는 추론을 하는 데 이용된다면 어떻게 될까? 시스템이 개인과 직접 상호작용하지 않을 때 동의를 얻을 수 있는가? 후자의 문제는 '타인의 사물인터넷'으로 명명되었다. 기업 환경 역시 권력의 불균형 문제를 제기하고 있다. 많은 직원들은 고용주가 자신의 개인정보(건강 관련 데이터 포함)를 어떻게 사용하는지에 대해 명확한 동의를 얻지 못하고 있다. 이러한 문제를 해결하기 위해 IEEE(2017)는 이러한 기업 문화를 청산하고 직원의 동의 없이 데이터가 수집되지 않도록 직원 데이터 영향 평가를 도입하도록 제안한다. 또한 데이터는 명시적으로 언급된 특정 목적에 대해서만 수집 및 사용되어야 하며, 최신 상태를 유지하고, 합법적으로 처리되어야 하며, 필요 이상으로 장기간 보관되어서는 안 된다고 주장한다. 피실험자가 데이터 수집 시스템과 직접적인 관계가 없는 경우 동의가 동적으로 이루어져야 하며, 수집 및 사용에 대한 데이터 선호와 한계를 해석하도록 설계

된 시스템이어야 한다.

IEEE는 "AI에 대한 인식과 이해를 높이려면 학부생과 대학원생들에게 AI와 지속 가능한 인간 발전과의 관계를 교육해야 한다"고 말한다. 이를 위해서는 구체적으로 커리큘럼과 핵심 역량이 정의되고 준비되어야 한다. 국제 개발 및 인도주의적 구제에 관한 공학에 초점을 맞춘 학위 프로그램은 AI 응용에 잠재적으로 노출되어야 한다. 그리고 전 세계의 인도주의적 노력에 있어서 AI의 구현을 통해 저소득 국가들이 직면할 수 있는 기회와 위험에 대한 인식을 높여야 한다.

책임로봇재단(Foundation for Responsible Robotics), AI파트너십(Parnership on AI), 일본AI윤리지침협회(Japanese Society for AI Ethical Guidelines), 미래사회협회(Future Society), AI현재연구소(the AI Now) 등 많은 협회들이 이 문제에 관심을 가지고 있다. 이들 등은 AI와 사회 간 명확하고 개방적이며 투명한 대화가 이해와 수용, 신뢰를 창출하기 위한 핵심이라는 점에 대해 의견을 고수하고 있다.

실존하는 위험

미래생명연구소(Future of Life Institute)에 따르면 AI를 둘러싼 주요 실존적 이슈는 '악의가 아니라 경쟁이다. AI가 타인과 교류하며 데이터를 수집하면서 지속적으로 학습해 시간이 지남에 따라 지능을 얻고 인간과 상충하는 목표를 설정할 수 있게 된다.

“당신은 악의로 개미를 밟는 사악한 개미 혐오자는 아닐 거예요. 하지만 당신은 수력 발전 녹색 에너지 프로젝트를 맡고 있고, 그 지역의 개미집은 물에 잠기게 되겠죠. 개미들에게는 너무 불쌍한 일이죠. AI 안전 연구의 핵심 목표는 절대 인간을 개미들의 위치에 두지 않는 것입니다(The Future of Life Institute, 2019).”

AI는 또한 자율 무기 시스템(AWS)의 형태로 위협을 제기한다. 이것들은 신체적인 해를 입히도록 고안되었기 때문에, 수많은 윤리적 문제를 일으킨다. IEEE(2019)는 AWS가 의미 있는 인적 통제를 받도록 하기 위한 여러 가지 권고 사항을 제시한다. 그들은 책임과 통제를 보장하기 위한 감사 추적, 투명하고 이해 가능한 방식으로 자신의 논리를 설명할 수 있는 적응형 학습 시스템을 제안한다. 그 시스템은 자신들의 작업물의 영향에 대해 알고 있어야 하고, 자율적 행동을 예측할 수 있어야 하고, 그리고 피해가 될법한 자율 시스템을 개발하는 데에 사용할 전문적인 윤리 규정을 마련해야 한다. AWS의 추구는 국제적인 군사력 경쟁과 지정학적 안정성을 초래할 수 있다. 그래서 IEEE는 인간의 통제나 판단의 경계 밖에서 행동하도록 설계된 시스템은 비윤리적이며 무기 사용에 대한 기본적인 인권과 법적 책임을 위반할 수 있다고 충고한다.

"사회를 심각하게 해칠 수 있는 잠재력을 감안할 때 이러한 문제는 미리 통제되고 규제되어야 한다"고 책임로봇재단(Foundation for Responsible Robotics)단은 주장한다. 이 위험을 명시적으로 다루는 다른 협회로는 국제노동조합네트워크(UNI Global Union)와 미래생명연구소(Future of Life Institute)가 있는데, 후자는 치명적인 자율 무기의 군비 경쟁에 대해 경고하고 가능성 있는 장기적 위험에 대한 계획 및 완화 노력을 요구한다. 우리는 미래 AI의 기능적 상한선에 대한 지나친 가정을 피해야 하고, FLI의 아실로마 원칙을 주장해야 하며, 진보된 AI가 지구 생명의 역사에서 중대한 변화를 나타낸다는 것을 인식해야 한다.

3.3 사례연구

3.3.1 사례연구: 의료 로봇

인공지능과 로봇공학은 의료분야에 빠르게 도입되고 있으며, 질병 진단과 임상 치료 분야에서 갈수록 더 활약할 것이다. 구체적인 예시로, 현재 혹은 가까운 미래에 로봇은 환자의 질병 진단, 간단한 수술 진행, 단기 및 장기 요양 시설에서 환자의 신체 및 정신 건강 모니터링을 도울 수 있을 것이다. 또한 로봇은 기본적인 신체적 개입도 가능하며, 가정에서 간병인의 역할도 수행할 수 있고, 환자에게 약 복용 시기를 알려줄 수 있으며, 거동이 불편한 환자도 도울 수 있을 것이다. 의료 영상 진단과 같은 의학의 일부 핵심 영역에서 인공지능의 질병 감지 능력은 인간과 같은 수준이거나 심지어 인간보다 뛰어난 것으로 드러났다.

구현된 인공지능(embodied AI) 혹은 로봇은 이미 사람들의 신체적 안전에 영향을 미치는 다양한 기능들에 관여하고 있다. 2005년 6월 필라델피아의 한 병원에서 수술용 로봇이 전립선 수술 도중 오작동하여 환자를 다치게 했다. 2015년 6월 독일 폭스바겐 공장에서 인부가 생산라인에서 로봇에 깔려 숨졌다. 2016년 6월에는 자율주행 모드로 주행하던 테슬라 승용차가 대형 트럭과 충돌해 승용차 안에 있던 탑승자가 사망하였다(Yadron and Tynan, 2016).

로봇이 점점 보편화 되면서, 로봇이 미래에 가져올 잠재적 위험 또한 증가하게 된다. 특히 무인자동차, 보조 로봇, 드론 등이 내리는 결정은 인간의 안전과 복지에 직접적인 영향을 미칠 것이다. 단순한 소프트웨어보다 구현된 인공지능, 로봇은 물리적 공간에서 움직이는 부품을 갖고 있으므로 로봇의 영향력은 더욱 클 것이다(Line et al.,2017). 움직이는 물리적 부품을 가진 로봇은 특히 어린이나 노약자와 같은 약자에게 위험을 초래할 수 있다.

안전

다시 말해, 의료 분야에서의 인공지능과 로봇공학이 발달로 인해 발생하는 가장 중요한 윤리적 문제는 안전과 상해 방지일 것이다. 로봇이 사람을 해치지 않고 함께 일하기에 안전해야 한다는 것은 매우 중요한 문제이다. 이 문제는 특히 질병이 있는 사람, 노인, 어린이와 같은 약자들을 다루는 의료 분야에서 중요하다. 디지털 의료 기술은 진단과 치료의 정확도 향상에 대한 가능성을 열어주지만, 기술의 장기적 안전성과 성능을 철저히 확립하기 위해서 임상시험에 대한 투자가 뒷받침되어야 한다. 질(음부) 메쉬 임플란트(골반 내장 기관 탈출증 / 요실금 교정 수술)로 인해 몸이 쇠약해지는 부작용과 메쉬 제조업체를 상대로 한 지속적인 법정 공방(The Washington Post, 2019)이 의료혁신을 늦출지라도 이는 성급하게 진행된 임상시험의 폐해를 보여준다. 인공지능 시스템이 제공하는 의료 혁신을 안전하게 구현하기 위해서는 임상시험에 대한 투자가 필수적일 것이다.

사용자 이해

환자의 안전을 보장하려면 의료 전문가가 인공지능을 올바르게 활용하는 것이 중요하다. 예를 들면, 정밀한 수술 보조 로봇 '다빈치'는 수술로 인한 상처를 최소화하는데 유용한 도구임을 입증받았지만, 숙련된 조작자가 필요하다(The Conversation, 2018).

의료 인력의 기술 균형에는 변화가 필요하며, 의료계는 향후 20년 동안 의료인들의 디지털 기술에 대한 이해 및 활용 능력을 개발할 준비를 하고 있다(NHS' Topol Review, 2009). 유전체학과 기계 학습(머신 러닝)이 진단과 의료 의사결정에 포함됨에 따라, 의료 전문가들은 각 기술적 도구를 이해하고 적절히 사용하기 위해 디지털 지식을 갖추어야 한다. 디지털 기술 사용자는 인공지능을 신뢰하되 각 도구의 장단점을 인식하여 검증이 필요한 시기를 인식하는 것이 중요하다. 예를 들어, 폐렴 환자의 합병증 위험성을 예측하는 기계 학습 연구는 합병증 위험성을 대체로 정확히 예측했지만, 천식 환자의 합병증 위험성은 낮다고 잘못 예측했다. 이러한 오류는 천식성 폐렴 환자들은 곧바로 중환자실로 옮겨져 집중 치료를 받음으로써 합병증을 피해 가기 때문에 발생하였다. 따라서 해당 알고리즘의 부정확한 권장 사항은 무시되었다(Pulmonology Advisor, 2017).

하지만 인공지능 시스템이 자율적이고 정보에 입각한 결정을 내리기 위해, 일정한 예측에 어떻게도달했는지를 인간이 어느 정도까지 이해하고 있어야 하는지는 의문이다. 수학에 대한 심층적인 이해가 의무화되더라도, 기계 학습 알고리즘의 복잡성과 학습된 특성으로 인해, 자료 집합(일명 '블랙박스')을 통한 결론 도출 과정은 보통 이해하기 매우 힘들다(Schönberger, 2019).

데이터 보호

의료 알고리즘에 활용되는 개인 의료 데이터는 위험에 처해 있을 수도 있다. 예를 들어, 피트니스 트래커에 의해 수집된 데이터가 보험 회사와 같은 제3자에게 판매되어 그 데이터를 토대로 의료 보험 가입을 거부할 수도 있다는 우려가 있다(National Public Radio, 2018). 해커 문제도 큰 우려 사항이다. 이에 대비하여 다양한 의료 인력이 접근하는 시스템에 대한 적절한 보안 시스템을 구축하는 것도 골칫거리일 것이다. (Forbes, 2018).

개인 의료 데이터를 한데 모으는 것은 기계 학습 알고리즘이 의료 개입을 진전시키는 데 중요하지만, 정보 거버넌스의 격차는 책임 있고 윤리적인 데이터 공유에 대한 장벽을 형성한다. 의료 전문가와 연구자가 유전체학과 같은 데이터를 환자들의 정보를 안전하게 지키면서 사용하는 방법에 대한 명확한 체계가 필요하다. 이를 통해 대중의 신뢰를 확립하고 의료 알고리즘을 발전시킬 수 있다. (NHS Topol Review, 2009).

법적 책임

인공지능은 의료사고를 발생 건수를 줄일 수 있겠지만, 만약 사고가 발생하면 법적 책임을 져야 한다. 장비의 결함이 입증될 수 있다면 제조사가 책임을 질 수 있지만, 시술 과정에서 무엇이 잘못됐는지, 수술자와 기계 중 누구에게 책임이 있는지 규명하기가 까다로운 경우가 많다. 예를 들어, 다빈치 수술 보조 로봇에 대한 소송이 있었지만(Mercury News, 2017), 로봇은 계속 널리 사용되고 있다(The Conversation, 2018).

결론 도출 과정 확인이 불가능한 '블랙박스' 알고리즘의 경우, 알고리즘 생산자 측의 과실을 규명하는 것은 상당히 까다롭다(Hart, 2018).

현재로서는 인공지능이 전문가의 의사결정을 위한 보조 역할을 하므로 대부분의 경우 전문가들이 책임을 지게 된다. 예를 들어, 앞서 언급한 폐렴 사례 같은 경우, 의료진이 오로지 인공지능에만 의존하여 자신의 전문 지식을 적용하지 않고 천식성 폐렴 환자를 집으로 돌려보냈다면, 그것은 의료진의 과실로 간주되었을 것이다(Pulmonology Advisor, 2017; International Journal of Law and Information Technology, 2019).

머지않아 인공지능을 사용하지 않은 것도 과실로 간주할 수 있을 것이다. 예를 들어, 의료 전문인력이 부족한 개발도상국에서 진단을 승인할 안과 의사가 부족하다고 해서 당뇨병성 안구질환을 발견하여 실명을 예방하는 인공지능의 도입을 보류하는 것은 오히려 비윤리적으로 여겨질 수 있다. (The Guardian, 2019; International Journal of Law and Information Technology, 2019).

편향

비차별은 EU의 핵심 가치 중 하나이지만(EU 기본권 헌장 21조 참조) 머신 러닝 알고리즘 훈련에 사용되는 데이터 세트는 소수자에 대한 데이터가 상대적으로 적기 때문에 편향될 수 있다(Medium, 2014). 이는 질환을 진단하기 위해 훈련된 알고리즘이 특정 인종의 환자를 정확하게 진단할 수 없을 수도 있다는 것을 의미한다. 예를 들어, 피부암을 감지 모델을 훈련하는 데 사용된 데이터 세트에서 유색인종의 이미지는 5% 미만에 불과해 유색인종을 상대로 오진할 위험성을 지닌다. (대서양, 2018)

모든 인종에게 정확한 진단이 제공되기 위해 알고리즘 편향을 식별하고 이해해야 한다. 모델 설계에 대한 완벽한 이해를 갖추고 있어도 앞에서 언급한 기계 학습의 '블랙박스' 특성 때문에 이는 어려운 작업이다. 하지만, 편향된 점을 발견하기 위해 다양한 지침이나 계획이 이미 도입된 바 있다. 예를 들어, 구글, 페이스북, 아마존, IBM, 마이크로소프트(The Guardian, 2016)가 협업하여 ‘인공지능 파트너십’이라는 윤리에 초점을 둔 산업 그룹을 개시하였지만, 걱정스럽게도 해당 그룹은 그리 다양성을 띠지 않는다.

접근의 평등

피트니스 트래커나 인슐린 펌프와 같은 디지털 건강 기술은 환자가 직접 자신의 건강관리에 적극적으로 참여할 기회를 제공한다. 어떤 사람들은 이러한 기술들이 열악한 교육, 실업 등으로 인한 건강 불평등을 해소하는 데 도움이 될 것이라고 믿는다. 다만 필요한 기술을 사용할 여력이 없거나 '디지털 능력'이 부족한 사람들은 배제돼 건강 불평등을 오히려 심화할 위험이 있다(The Guardian, 2019).

영국 국립 보건 서비스의 디지털 참여 확대 프로그램은 건강 불평등을 완화하기 위해 시행되었다. 해당 프로그램은 디지털 의료 서비스에 접근할 수 있는 기술이 부족한 영국 내 수백만 명의 사람들을 도와주었다. 이와 같은 의료에 대한 접근의 동등성을 보장하면서도 위에서 논의한 의료 알고리즘의 편향을 방지하기 위해 필요한 소수 그룹의 데이터를 늘릴 수 있는 프로그램은 매우 중요하다.

진료의 질

'디지털 의료 기술에는 진단 및 치료의 정확성, 진료의 효율성 및 의료 전문가의 작업 흐름을 개선할 수 있는 놀라운 잠재력이 있다.'(NHS 'Topol Review, 2019).

동반자 및 돌봄 로봇이 신중한 지침을 갖춰 도입된다면, 노인들의 삶을 개선하고, 노인들의 타인 의존도를 낮추고, 더 많은 사회적 상호 작용 기회를 제공할 것이다. 우리는 약 복용 시기를 알려주고, 피곤하거나 이미 침대에 누워 있는 사람을 위해 물건을 가져오고, 간단한 청소를 대신해 주거나 영상 통화로 가족, 친구나 의료 서비스 제공자와 연락을 유지할 수 있도록 도와주는 가정용 돌봄 로봇을 상상해 볼 수 있다. 하지만, '차갑고 감정 없는' 로봇이 과연 인간의 감정적인 손길을 대체할 수 있을지에 대한 의문이 제기된다. 특히 장기적으로 간호가 필요한, 간병인과 기본적인 우정을 쌓는 취약하고 고독한 환자들이 이에 해당한다. 연구에 의하면 특히 인간의 상호작용은 나이가 들수록 중요해지는데, 광범위한 인간관계가 치매를 예방할 수도 있기 때문이다. 현재로선 로봇은 진정한 동반자와는 거리가 멀다. 비록 로봇이 사람들과 교류할 수 있고, 심지어 가짜여도 감정을 표현할 수 있지만, 대화 능력이 여전히 극도로 제한적이며, 인간의 사랑과 관심을 대체하지는 못한다. 어떤 사람들은 노인들의 인간 접촉을 박탈하는 것은 비윤리적이며, 심지어 폭력의 한 형태라고도 말할 수 있다.

노인들을 차가운 기계가 돌보도록 놔두는 것이 누구를(의) 물건 취급(존엄성을 저하)하는 것인가? 노인들인가? 인간 간병인인가? 로봇 때문에 노인들이 자신들이 물건 취급을 당하고 있다고 느끼지 않게 하고, 인간 간병인에게 의지했을 때보다 자신들의 삶에 대한 통제력이 더 떨어졌다고 느끼지 않게 하는 것은 매우 중요하다. 그렇지 않으면 노인들은 '지각이 있는 존재라는 언급이 없이는 자신들이 죽은 물질 덩어리에 지나지 않아 밀쳐지고, 들어 올려지고, 펌프질 되거나, 배출되는 것'이라고 느낄 것이다. (Kitwood, 1997)

이론상으로는 자율성, 존엄 및 자기 결정권 모두 기계적 응용으로 충분히 준수될 수 있지만 민감한 의학 분야에서 이런 역할의 응용이 허용될 수 있을지는 미지수다. 예를 들어, 한 의사가 원격회의를 통해 캘리포니아 환자에게 사망 예후를 알려 주었다. 아니나 다를까 환자 가족은 이러한 인간미 없는 의료 접근 방식에 분노했다(The Independent, 2019). 한편, 건강 모니터링 앱과 같은 신기술은 의료진에게 환자와 더욱 직접적인 상호작용을 위한 시간을 확보해 주어 전반적인 진료의 질을 높일 수 있다는 주장이 있다(The Guardian, Press Association, 2019년 2월 11일 월요일).

기만

수많은 '케어봇'이 사회적 상호작용을 위해 설계되었으며, 감정 치료 역할도 제공한다고 광고한다. 예를 들어, 요양원에서 아기 물개 로봇의 동물을 닮은 상호작용이 환자들의 기분을 좋게 하고, 불안감을 줄이며, 환자들과 간병인 간의 친화력을 증가시킨다는 것을 발견했다. 하지만 치매 환자에게는 현실과 상상의 경계가 모호하기 때문에 로봇을 반려동물이라 소개하여 사회 정서적 개입을 조장하는 것이 과연 부정한 행위인가? (KALW, 2015) 만약 그렇다면 도덕적으로 정당화될 수 있는가?

동반자 로봇과 애완동물 로봇은 노인들의 외로움을 덜어줄 수 있지만, 로봇이 노인들을 배

려하고 감정을 가진 지각 있는 존재라고 믿게끔 해야만 가능하다. 근본적인 기만이라 할 수 있다. Turkle 등(2006)은 '우리 부모, 조부모, 아이들이 '사랑해'라고 대답해줄 로봇에게 '사랑해'라고 말할 수도 있다는 사실이 편하지는 않다. 우리가 요구하는 기술의 진정성에 의문을 제기한다'고 주장한다. Wallach와 Allen(2009)은 인간의 사회적 제스처를 감지하고 이에 동일하게 반응하도록 설계된 로봇 모두 기만의 형태인 기술을 사용한다는 것에 동의한다. 인간이 로봇 애완동물을 통해 이익을 얻으려면, 실제 동물과의 관계를 맺고 있다고 자기 자신을 속여야 한다. 게다가, 노인들에게 로봇 장난감과 상호작용하도록 장려하는 것은 노인들을 어린아이로 취급하는 것과 같다.

자율성

중요한 점은 의료 로봇이 실제로 환자에게 도움이 된다는 것이고, 그 밖의 사회 의료 부담을 줄이는 것만을 목적으로 하지 않는다는 점이다. 특히 돌봄 및 동반자 인공 지능의 경우 더욱 더 그렇다. 로봇은 장애인이나 노인에게 권한을 부여하고 그들의 독립성을 높일 수 있다. 실제로 일부 장애인이나 노인은 화장실이나 목욕과 같은 특정 사적인 일 같은 경우, 인간의 도움보다 로봇을 선호할 수 있다. 로봇은 노인들이 자신의 집에서 더 오래 살 수 있도록 도와주어 더 큰 자유와 자율성을 줄 수 있다. 하지만 만약 정신적 능력에 문제가 있는 사람에겐 얼마만큼의 통제 혹은 자율성이 허용되어야 하는가? 만약 환자가 로봇에게 자신을 발코니에서 던져 달라고 요청하면 로봇은 그 명령을 수행해야 하는가?

자유와 사생활

인공지능 기술의 여러 분야와 마찬가지로 의료 서비스와 동반자 로봇을 설계할 때에도 이용자의 사생활과 존엄성을 신중하게 고려할 필요가 있다. 사람들의 집에서 일하는 것은 로봇이 목욕이나 옷 입는 것과 같은 사적인 순간에 관여할 수 있음을 의미한다. 만약 이러한 순간이 기록된다면, 누가 이 정보에 접근할 권한이 있고, 그 기록은 얼마나 오랫동안 보관되어야 하는가? 노인의 경우, 정신 상태가 악화되어 혼란을 겪게 되면 문제는 더욱 복잡해진다. 알츠하이머를 앓고 있는 사람은 로봇이 자신을 감시하고 있었다는 사실을 잊어버릴 수 있어 집에 자신만 있다고 착각한 상태로 행동이나 말을 할 수 있다. 홈케어 로봇은 의료상 위급한 경우를 제외하고 환자의 방에 들어가기 전에 노크하고 초대를 기다리는 행위 등을 통해 사용자의 사생활과 간호가 필요한 상황 간의 균형을 맞출 수 있어야 한다.

로봇은 환자의 안전을 위해 때때로 자유를 제한하고 관리자 역할을 해야 할 수도 있다. 예를 들어, 가스레인지가 켜져 있거나 욕조가 넘칠 경우 로봇이 개입하도록 훈련될 수 있다. 로봇은 심지어 노인들이 찬장에서 무언가를 꺼내기 위해 의자 위로 올라가는 것과 같은 잠재적으로 위험한 행동을 하는 것을 막아야 할 수도 있다. 센서가 달린 스마트 홈은 한 사람이 방을 나가려고 하는 것을 감지하고 문을 잠그거나 의료진을 부르는 데 사용될 수 있지만, 그럴 경우 환자는 방에 갇히게 될 것이다.

도덕적 기능

'뇌를 활용해 사물을 제어하려는 흥미로운 연구가 있습니다. 아마 팔을 잃어버린 사람들을 위한 일이겠죠… 제일 우려되는 부분은 행동 타게팅과 같은 문제입니다. 지금처럼 데이터 접근을 위해 사람들이 '동의'를 누르듯이 해마로 직행하여 '동의'할 것입니다.' (John Havens)

로봇은 윤리적 성찰 능력이나 의사결정을 위한 도덕적 기반이 없어 현재로선 인간이 의사결정에 대한 궁극적인 통제권을 가지고 있어야 한다. 로봇의 윤리적 추론의 예는 2004년 디스토피아 영화 '아이, 로봇'에서 찾아볼 수 있다. 윌 스미스가 연기한 등장인물은 가상 미래 시대의 로봇들이 냉정한 논리 판단으로 아이의 생명 대신 자신의 생명을 구하는 것을 반대했다. 더 많은 자동화된 의료 서비스가 도입된다면 도덕적 기능과 관련된 문제는 더 면밀한 관심이 필요할 것이다. 윤리적 추론이 로봇에 탑재되어가고 있지만, 도덕적 책임은 윤리의 적용보다 더 많은 것을 의미한다. 미래의 로봇이 의료분야에서 발생하는 복잡한 도덕적 문제를 다룰 수 있을지는 불확실하다(Goldhill, 2016).

신뢰

라로사(Larosa)와 당크스(Danks)(2018)는 인공지능이 의료 영역 내에서 인간과 인간 간의 상호 작용, 특히 환자와 의사 간의 관계에 영향을 미칠 수 있으며, 잠재적으로 의사에 대한 신뢰를 무너뜨릴 수 있다고 한다.

'심리학 연구에 의하면, 사람들은 컴퓨터와 같이 비용과 이익을 계산함으로써 도덕적 결정을 내리는 사람들을 불신하는 경향을 보인다고 한다.' (The Guardian, 2017) 로봇에 대한 우리의 불신은 디스토피아 공상 과학 소설에서 날뛰는 로봇의 수에서 발생할 수도 있다. 컴퓨터의 오류에 관한 뉴스 기사들을 예로 들자면, 이미지 식별 알고리즘이 거북이를 총으로 오인하는 경우(The Verge, 2017)가 발생했다. 이처럼 알려지지 않은 것이나 프라이버시 및 안전성에 대한 우려들이 인공지능의 도입을 거부하는 이유가 된다(Global News Canada, 2016).

첫째, 의사는 명시적으로 의료행위를 할 수 있는 자격증과 면허가 있으며, 이러한 자격은 의사가 '해를 끼치지 않는다'라는 특정한 기술, 지식, 가치를 지녔음을 의미한다. 로봇이 특정 치료나 진단 작업을 위해 의사를 대체하는 경우, 환자는 이 로봇 시스템이 수행하는 기능이 적절히 승인되거나 '면허가 있는지'의 여부를 알아야 한다. 이 때문에 로봇은 잠재적으로 환자와 의사 사이의 신뢰를 위협할 수 있다.

둘째로, 환자는 의사가 전문 지식의 모범이라고 생각하기 때문에 신뢰한다. 의사들이 인공지능의 '단순한 사용자'로 인식된다면, 대중의 눈에 의사의 역할이 격하되어 그들에 대한 신뢰가 떨어질 것이다.

셋째로, 환자가 의사와 함께한 경험은 중요한 신뢰의 원동력이다. 환자가 의사와 활발한 의사소통을 할 수 있고, 진료와 치료에 대한 대화를 나눈다면 환자는 의사를 신뢰할 것이다. 반대로, 의사가 환자의 요청을 반복적으로 무시한다면, 의사의 신뢰에 부정적인 영향을 미칠 것이다. 이러한 상황에 인공지능을 도입하면 신뢰를 높일 수 있다. 인공지능이 오진의 가능성을 줄이거나 환자 관리 능력을 키운다면 말이다. 하지만 인공지능은 오히려 의사의 신뢰를 떨어

트릴 수도 있다. 의사가 인공지능에 의료 문제에 대한 권한 혹은 진단이나 의사 결정 권한을 지나치게 많이 위임하면 의사 자신의 의료 분야 권위자로서의 입지가 축소된다.

각 기술 접근법에 대한 치료 이점을 지원하는 증거가 많아지고 로봇 상호 작용 시스템의 시장 유입이 늘어남에 따라 로봇에 대한 신뢰가 높아질 것으로 보인다. 이는 다빈치 수술 보조 로봇과 같은 로봇 건강 관리 시스템에서 이미 발생한 바 있다(The Guardian, 2014).

고용 대체

다른 산업과 마찬가지로 신흥 기술의 등장은 고용을 위협할 우려가 있다(The Guardian, 2017). 그 예시로, 간호사가 수행하는 업무의 3분의 1을 대신 담당할 수 있는 의료 로봇들이 있다(Tech Times, 2018). 이러한 우려에도 불구하고, NHS' Topol Review (2009)는 '이런 기술이 의료 전문가들을 대체하는 것이 아니라 그들의 업무 능력이 향상되어 환자에게 쏟을 시간을 더 늘릴 수 있다'고 결론 내렸다. 해당 보고서는 영국 NHS가 직원들의 디지털 능력을 키우기 위해 학습 환경을 어떻게 조성할 것인지에 대한 밑그림을 제공했다.

3.3.2 사례 연구 : 자율주행차량

자율 주행 차량(AV)은 주위의 환경을 감지할 수 있고 운전자가 거의 또는 전혀 조작하지 않아도 되는 차량이다. 자율주행차에 대한 아이디어는 적어도 1920년대부터 있었지만, 공공 차로에 등장할 정도로 기술이 발달한 것은 최근 몇 년이 채 되지 않았다.

자동차 표준화 기구 SAE(미국자동차공학회)(2018)는 자율주행 단계를 6단계로 발표했다.

0	비자동화	자동 시스템이 경고를 하거나 순간적으로 주행에 개입할 수 있지만 지속적인 차량 제어는 하지 않는다.
1	운전자 보조	운전자와 자동 시스템이 차량에 대한 제어권을 공유한다. 예를 들어, 자동 시스템은 설정된 속도(예: 크루즈 컨트롤), 속도를 유지 및 변경하기 위한 엔진 및 브레이크 동력(예: 어댑티브 크루즈 컨트롤) 또는 스티어링(예: 주차 지원)을 제어할 수 있다. 운전자는 언제든지 모든 제어를 할 준비가 되어 있어야 한다.
2	부분 자동화	자동 시스템이 차량(가속, 제동 및 스티어링 포함)을 완전히 제어한다. 단, 운전자는 운전을 모니터링하고 언제든지 즉시 개입할 수 있도록 준비해야 한다.
3	조건부 자동화	차량이 즉각적인 응답을 요구하는 모든 상황을 처리할 것이기 때문에 운전에 신경을 쓰지 않아도 된다(예: 텍스트 또는 필름 작업). 그러나, 운전자는 시스템이 지정한 시간 내에, 시스템에 의해 그렇게 하도록 요청 받은 경우에도 개입할 준비가 되어 있어야 한다.
4	고도 자동화	레벨3과 비슷하지만, 안전을 위해 운전자의 주의가 필요하지 않으므로 운전자가 잠을 자거나 운전석에 있지 않아도 된다.
5	완전 자동화	사람의 개입이 전혀 필요하지 않다. 레벨5의 예시로는 로봇 택시가 있다.

하위 레벨의 자동화 중 일부는 이미 잘 확립되어 있고 시장에 출시되어 있으며, 상위 레벨의 자율주행 기술은 개발 및 테스트를 거치고 있다. 그러나, 우리가 자동화 레벨을 높이고 인간 운전자보다 자동화 시스템에 더 책임을 많이 지우게 되면서, 많은 윤리적 문제가 대두되고 있다.

자율주행의 사회적, 윤리적 영향

'우리는 "인간이 특정한 방식으로 행동한다는 것을 알고 있다. 우리는 그들을 죽일 것이다"라고 말하는 도구를 만들 수 없다.(John Havens)'

공공도로에서 테스트되는 공공안전과 윤리

현재 대부분의 국가에서 '보조 운전' 기능이 있는 자동차는 합법적이다. 특히 일부 테슬라 모델에는 레벨 2 자동화(Tesla, nd)를 제공하는 오토파일럿 기능이 있다. 운전자는 항상 차량을 책임지고 있는 경우에 공공 도로에서 보조 운전 기능을 사용할 수 있다. 그러나 이러한 보조 운전 기능 중 많은 부분이 아직 독립적인 안전 인증을 받지 않았으며, 따라서 운전자와 다른 도로 사용자에게 위험이 될 수 있다. 독일에서는 자동운전윤리위원회가 발간한 보고서에서 공공도로에 도입되어 허가를 받은 자율주행 시스템의 안전성을 보장하는 것이 공공부문의 책임임을 강조하고, 모든 자율주행 시스템은 공식 인증 및 모니터링 대상이 되어야 한다는 것을 권고하고 있다(Ethics Commision, 2017).

아울러 자동차가 아직 완전히 자율화되지 않았지만 인간 운전자가 제대로 관여하지 않는 등 자율주행차 산업이 가장 위험한 국면으로 접어들고 있다는 의견이 제기됐다(Solon, 2018). 이러한 위험은 자율주행차가 관련된 최초의 보행자 사망 이후 널리 주목을 받았다. 비극은 2018년 5월 미국 애리조나에서 우버(Uber)의 테스트를 받던 레벨 3 자율주행차가 어느 날 밤 자전거를 타고 길을 건너던 49세의 일레인 허즈버그(Elaine Herzberg)와 충돌하면서 발생했다. 검찰은 우버가 '형사상 책임이 없다'(Shepherdson and Somerville,2019)고 판단했고, 원인에 대한 결론을 도출하지 못한 미 교통안전위원회 예비보고서(NTSB·2018)는 충돌 당시 자율주행 시스템의 모든 요소가 정상적으로 작동했다고 밝혔다. 우버는 운전자가 비상 제동이 필요한 상황에서 개입하고 조치를 취하는 데에 의존한다고 말했으며, 이에 따라 일부 해설자는 '자율주행차'와 '자동 조종차'라는 용어에 대해 소비자들에게 오해를 불러일으키기도 했다(Legget, 2018). 이 사고로 일각에서는 공공도로에서 자율주행차 시스템을 시험하는 관행이 위험하고 비윤리적이라고 비난했고, 우버는 자율주행 프로그램을 잠정 중단했다(Bradshaw, 2018).

사람의 안전 문제는(공공안전과 승객안전 모두) 자율주행 자동차와 관련된 핵심 이슈로 떠오르고 있다. 닛산, 도요타, 테슬라, 우버, 폭스바겐 등 주요 기업들은 인간의 직접적인 통제 없이 복잡하고 예측 불가능한 환경에서 운전할 수 있고 학습, 추론, 계획 및 의사결정이 가능한 자율 주행 차량을 개발하고 있다.

자율주행차는 여러 가지 이점을 제공할 수 있다. 통계에 따르면 사람이 운전하는 차보다 컴퓨터가 운전하는 차가 더 안전합니다. 또한 도시의 혼잡함을 완화하고, 공해를 줄이고, 여행과 통근 시간을 줄이고, 사람들이 시간을 더욱 생산적으로 사용할 수 있도록 할 수 있다. 하지만, 그들이 도로 교통 사고를 내지 않는다는 것을 의미하지는 않는다. 자율주행차가 최고의 소프트웨어와 하드웨어를 갖췄다고 해도 충돌 위험은 여전히 있다. 예를 들어 주차된 차량 뒤에서 튀어나온 어린이에 의해 깜짝 놀랄 수 있으며, 이러한 자동차가 누구의 안전을 우선시해야 할

지를 결정해야 할 때 어떻게 프로그래밍되어야 하는지에 대한 문제가 항상 존재한다.

운전자 없는 자동차도 탑승자과 다른 보행자의 안전 중 하나를 선택해야 할 수도 있다. 자동차가 한 무리의 학생들이 놀고 있는 모퉁이를 돌면, 차를 멈추게 할 시간이 충분치 않으며, 차가 아이들을 치지 않을 수 있는 유일한 방법은 벽돌담을 들이받아 탑승자를 위험에 빠뜨리는 것 뿐이다. 어린이와 탑승자 중 누구의 안전을 최우선으로 해야 하는가?

사고 조사를 위한 프로세스 및 기술

자율주행차는 고급 머신러닝 기술에만 의존하는 복잡한 시스템이다. 이미 레벨 2 자율주행차와 관련된 다수의 사망자를 포함하여 몇 가지 심각한 사고가 발생했다.

▷ 2016년 1월, 23세 가오야닝(Gao Yaning)은 중국 허베이시 고속도로에서 자신의 테슬라 모델 S가 도로 청소용 트럭 뒷부분을 들이받아 사망했다. 가족들은 오토파일럿이 사고가 났을 때 관여한 것으로 보고 테슬라가 시스템 기능을 과대평가 했다고 비난했다. 테슬라는 차량 파손으로 오토파일럿이 작동했는지, 오작동했는지의 여부를 판단할 수 없었다고 밝혔다. 충돌에 대한 민사 소송이 진행 중이며, 제3자인 평가자가 차량의 데이터를 검토하고 있다(Curtis, 2016).

▷ 2016년 5월 미국 플로리다에서 오토파일럿이 작동하던 중 테슬라 모델 S가 트럭과 충돌해 사망한 조슈아 브라운(Joshua Brown)은 미국도로교통안전청 조사 결과 테슬라가 아닌 운전자의 과실로 밝혀졌다(Gibbs, 2016). 그러나, 미국도로교통안전국은 후에 오토파일럿과 테슬라의 운전 보조 장치에 대한 운전자의 과의존에 모두 책임이 있다고 판단하였다(Felton, 2017).

▷ 웨이황(Wei Huang)은 2018년 3월 미국 캘리포니아에서 자신의 테슬라 모델 X가 고속도로 안전장벽을 들이받아 사망했는데, 당시 사고의 심각성은 '전례가 없는' 수준이었다고 테슬라는 전했다. 미국 교통안전위원회는 후에 이 사고가 오토파일럿 항법 실수 때문이라고 하는 보고서를 발표했다. 테슬라는 현재 피해자 가족과 소송을 진행중에 있다.(O'Kane, 2018)

불행히도, 이러한 사고를 조사하기 위한 노력은 자율주행차와 관련된 사고를 조사하기 위한 표준 프로세스 및 규제 프레임워크가 아직 개발되거나 채택되지 않았기 때문에 좌절되었다. 또한, 현재 자율주행차에 설치되어 있는 독점 데이터 로깅 시스템은 사고 조사자가 사고로 이어지는 사건에 대한 중요한 데이터를 제공하기 위해 제조업체의 협력에 크게 의존하고 있음을 의미한다(Stillgoe and Winfield, 2018).

한 가지 해결책은 미래의 모든 자율주행차에 독립적인 사고 조사자가 액세스할 수 있는 업계 표준 이벤트 데이터 기록 장치(일명 '윤리 블랙박스')를 장착하는 것입니다. 이는 항공 사고 조사를 위해 이미 시행 중인 블랙박스를 모델로 반영한 것이다(Sample, 2017).

하마터면 놓칠 뻔한 사고

현재, 거의 미수에 가까운 사고의 체계적인 수집을 위한 시스템은 마련되어 있지 않다. 제조사가 이미 사고 데이터를 수집하고 있을 수는 있지만, 그렇게 해야 하거나 데이터를 공유할 의무는 없다. 현재 유일한 예외는 미국 캘리포니아 주인데, 공공 도로에서 자율주행차를 적극적으로 테스트하고 있는 모든 회사는 안전상의 이유로 인간 운전자가 차량을 통제한 빈도(일명 '자율주행모드 해제')를 공개해야 한다.

2018년 자율주행차 제조사에 따른 자율주행모드 해제 횟수는 웨이모의 11,017마일마다 1회씩에서 애플의 1.15마일마다 1회씩까지 매우 다양했다(Hawkins, 2019). 이러한 해제에 관한 데이터는 인간 운전자들이 계속 관여하고 있는지 확인해야 하는 중요성을 강화한다. 그러나

일부에서는 캘리포니아의 데이터 수집 프로세스가 표현이 모호하고 엄격한 지침이 없기 때문에 기업들이 실수라고 주장하는 특정 사건들의 보고를 피할 수 있다고 주장하면서 비판을 받아왔다.

이러한 유형의 데이터에 접근하지 않으면 정책 입안자는 간과할 수 있는 사고의 빈도와 중요성을 설명할 수 없으며, 거의 놓칠 뻔한 결과로 제조업체가 취한 조치들을 평가할 수 없다. 다시 말하지만, 항공 사고 조사에 이은 모델로부터 교훈을 얻을 수 있었고 그 덕분에 모든 사고에 근접한 사건들이 철저히 기록되고 독립적으로 조사된다. 정책 입안자들은 규제를 알리기 위해 모든 사고와 하마터면 놓칠 뻔한 사고에 대한 종합적인 통계를 요구한다.

데이터 보안

제조업체가 자율주행차로부터 상당한 양의 데이터를 수집하는 것이 분명해지고 있다. 이러한 차량이 도로에서 점점 더 보편화됨에 따라, 이러한 데이터가 운전자와 승객의 개인 정보 보호 및 데이터 보호 권한을 어느 정도까지 훼손하고 있는가 하는 의문이 대두되고 있다.

이미 데이터 관리 및 개인 정보 보호 문제가 나타났으며, 일각에서는 자율주행 데이터를 광고 목적으로 오용할 수 있다는 우려를 제기하고 있다(Lin, 2014). 테슬라는 자율주행 데이터 로그의 비윤리적인 사용에 대해서도 비난을 받고 있다. 신문은 가디언 조사 결과 자사 기술이 사고에 영향을 미치지 않는다는 것을 입증하기 위해 무단으로 충돌 후 운전자의 개인 데이터를 언론과 공유한 사례가 다수 발견됐다(Thielman, 2017). 동시에 테슬라는 고객이 자신의 데이터 로그를 볼 수 있도록 허용하지 않는다.

독일 자동 운전 윤리 위원회가 제안한 한 가지 해결책은 모든 자율주행차 운전자에게 완전한 데이터 주권이 주어지도록 하는 것이다(Ethics Commission,2017). 이렇게 하면 운전자의 데이터가 어떻게 사용되는지 제어할 수 있다.

고용문제

자율주행차의 성장은 버스, 택시, 트럭 운전사 등 특정 직업군을 어려움에 빠뜨릴 가능성이 높다.

중기적으로는 장거리 트럭이 자율주행 기술의 최전선에 있기 때문에 트럭 운전자들이 가장 큰 위험에 직면하게 될 것이다(Viscelli, 2018). 2016년, 최초로 인간의 행동이 수반되지 않은 채로 맥주를 상업적으로 배달하기 위해 120마일을 주행하는 했다(Isaac, 2016). 지난해 자율주행 트럭이 처음으로 완전히 운전자 없는 운전을 했는데, 사람이 한 명도 타지 않은 채로 7마일을 주행했다(Cannon, 2018).

미래를 내다보면, 점점 더 많은 버스 운전사들이 운전하지 못하게 되고 일자리를 잃을 가능성이 높다. 에든버러(Calder, 2018), 뉴욕(BBC, 2019), 싱가포르(BBC, 2017) 등 전 세계 수많은 도시들이 자율주행 셔틀 도입 계획을 발표했다. 라스베이거스 셔틀은 운행 첫날 충돌사고로 출발이 험난했던 것으로 유명해졌고(Park, 2017), 스위스의 작은 마을 뉴하우젠 라인폴의 관광객들은 이제 자율주행 버스를 타고 인근 폭포를 방문할 수 있다(CNN, 2018). 중기적으로 운전자 없는 버스는 100% 전용 버스 차선을 따라 이동하는 노선으로 제한될 가능성이 높다. 그럼에도 불구하고, 자율주행 셔틀의 발전은 이미 미국의 노조와 시 공무원들에게 긴장을 조성했다(Weinberg, 2019). 지난해 미국 교통노동조합(Transport Workers Union of America)은 오하이오주(Pfleger, 2018)에서 자율버스가 부딪히는 것을 막기 위해 협정을 결성했다.

완전 자율 택시는 자율주행 기술이 완전히 테스트되고 레벨 4, 5에서 입증되면 먼 미래에 현실화될 가능성이 높다. 그럼에도 2021년까지 런던에서 자율주행 택시를 도입할 계획이 있고(BBC, 2018), 미국 애리조나에서 이미 자율 택시 서비스를 이용할 수 있는 상황에서 택시기사들이 불안해하는 이유를 쉽게 알 수 있다(Sage, 2019).

도시환경의 품질

장기적으로 자율주행차는 우리의 도시 환경을 재편성할 잠재력을 가지고 있다. 이러한 변화들 중 일부는 보행자들, 자전거 타는 사람들, 그리고 지역 주민들에게 부정적인 결과를 가져올 수 있다. 운전이 더욱 자동화됨에 따라 추가 인프라(예: 자율주행 전용 차선)가 필요할 가능성이 높다. 또한 자동화가 교통 혼잡과 주차에서부터 녹지 공간 및 로비에 이르기까지 모든 것을 구체적으로 계획함으로써 도시 계획에 더욱 광범위한 영향을 미칠 수 있다(Marshall and Davies, 2018). 자율주행차를 출시하려면 5G 네트워크의 범위가 크게 확장되어야 한다. 다시 말하지만, 확실히 도시 계획에 시사하는 바가 있는 것이다. (Kosravi, 2018).

자율주행차가 환경에 미치는 영향도 고려해야 한다. 자율주행차는 연료 사용량 및 관련 배출량을 크게 줄일 수 있는 잠재력이 있지만, 자율주행차는 장거리 주행이 편리하고 매력적이라는 점에서 이러한 절감 효과가 반작용할 수 있다(World, 2016). 따라서 자동화가 운전 행동에 미치는 영향을 과소평가해서는 안 된다.

법적 및 윤리적 책임

법적인 관점에서, 로봇에 의해 야기된 충돌은 누가 책임지고, 알고리즘에 의해 제어된 차량 때문에 부상을 입었을 때 피해자는 어떻게 보상받아야 하는가? 법원이 이 문제를 해결하지 못하면 로봇 제조사들은 예상치 못한 보상 비용을 초래해 투자를 위축시킬 수 있다. 다만, 피해자가 제대로 보상받지 못할 경우 자율주행차는 일반인이 신뢰하거나 받아들이지 않을 것으로 보인다.

로봇은 불확실하거나 '이길 수 없는' 상황에서 판단을 내려야 할 것이다. 하지만, 법률 지침이 없을 때 로봇이 따라야 할 윤리적인 접근법이나 이론은 무엇인가? 라인 등 여러 사람들이 설명하듯이, 다른 접근 방식은 충돌 사망자의 수를 포함하여 또 다른 결과를 만들어낼 수 있다.

또한, 운전자, 소비자, 승객, 제조업체, 정치인 등의 사람들 중에서 자율 주행 차량의 윤리는 누가 선택해야 하는가? 로와 로(Loh and Loh)는 엔지니어, 운전자, 자율주행 시스템 자체에서 책임을 분담해야 한다고 주장한다(2017). 그러나 밀러(Millar)는 이 경우엔 기술의 사용자, 즉 자율주행차 탑승자가 로봇이 따라야 할 윤리 또는 행동 원칙을 결정할 수 있어야 한다고 주장한다(2016). 의사들은 환자의 사전 동의 없이 연명치료에 대한 중요한 결정을 내릴 도덕적 권한이 없다는 것을 예시로 들어 엔지니어가 운전자에게 직접 의견을 묻거나, 특정 상황에서 차가 어떻게 행동할지에 대한 알고리즘을 사용자에게 미리 알리지 않고 차를 설계하면 도의적 반발이 있을 것이라고 주장한다.

3.3.3 사례연구: 전쟁과 무기화

2차 세계대전 이후 군사기술에 부분적으로 자율적이고 지능적인 시스템이 사용됐지만, 머신러닝과 AI의 진보는 전쟁에서의 자동화 도입의 전환점을 의미한다.

AI는 이미 위성영상분석과 사이버방어 등 분야에서 활용될 만큼 고도화돼 있지만, 진정한

적용범위는 아직 완전히 실현되지 않았다. 최근의 한 보고서는 AI 기술이 핵무기, 항공기, 컴퓨터, 생명공학 기술의 출현과 같거나 어쩌면 더 큰 규모로 전쟁을 변화시킬 수 있는 잠재력을 가지고 있다고 결론지었다(Allen and Chan, 2017). AI가 군에 영향을 미칠 수 있는 몇 가지 주요 방법이 아래에 정리되어 있다.

살상 자율 무기

자동적이고 자율적인 시스템들이 점점 더 능력을 발휘함에 따라, 군 당국은 자동화 시스템에 권한을 많이 위임하려고 하고 있다. 이는 AI의 광범위한 채택과 함께 계속될 것으로 보이며, AI에 영감을 받은 군비 경쟁으로 이어질 것으로 보인다. 러시아 군사 산업 위원회는 러시아 전투력의 30%를 2030년까지 완전히 원격 조종과 자율적인 로봇 플랫폼으로 구성하겠다는 공격적인 계획을 이미 승인했다. 다른 나라들도 비슷한 목표를 세울 것이다. 미국 국방부가 살상력 있는 자율 및 반자율 시스템의 사용을 제한한 반면, 다른 국가와 비국가 행위자들은 그러한 제한을 하지 않을 수도 있다.

드론 기술

보통의 군용기를 사려면 대당 1억 달러 이상의 비용이 들 수 있지만, 고품질 쿼드콥터 무인항공기는 현재 약 1,000달러이며, 이는 한 대의 고급 항공기 가격에 백만 대의 드론을 구입할 수 있다는 것을 의미한다. 현재 상용 드론은 사거리가 제한적이지만 미래에는 탄도미사일과 비슷한 사거리를 보유할 수 있어 기존 플랫폼을 쓸모없게 만들 수 있다.

로봇 암살

저비용, 고효율, 치명적, 자율적 로봇의 광범위한 가용성은 표적 암살을 더욱 만연하게 만들 수 있다. 자동 저격 로봇은 멀리서 목표물을 암살할 수 있다.

무선로봇폭발장치

상용 로봇 및 자율 주행 차량 기술이 널리 보급됨에 따라 일부는 이를 활용하여 보다 진보된 IED(Inimprovised Volutionary Device)를 만들 것이다. 현재, 수마일 떨어진 곳에서부터 정확한 목표지점에 폭발물을 신속하게 운반할 수 있는 기술적 능력을 가지고 있는 것은 강대국으로 제한되어 있다. 하지만 무인기에 의한 장거리 화물 배송이 현실화된다면, 폭발물을 정확하게 운반하는 비용은 수백만 달러에서 수천, 심지어 수백 달러로 떨어질 것이다. 마찬가지로, 자율주행차는 더 이상 자살을 각오한 운전자를 필요로 하지 않기 때문에 차를 이용한 자살 폭탄이 더 빈번하고 파괴적으로 될 수 있다.

한라크(Hallaq) 외 연구진은 또한 머신러닝이 전쟁에 영향을 미칠 가능성이 있는 주요 영역을 강조한다(2017). 이들은 CO(Command Officer)가 위성 이미지를 자동으로 스캔하여 특정 차량 유형을 감지하여 위협을 사전에 식별하는 데 도움을 주는 유동 전장 환경 내에서 지능형 가상 보조 장치(IVA)를 사용할 수 있는 예시를 보여준다. 또한 적의 의도를 예측하고, 상황 데이터를 수백 개의 이전 전쟁 모의 연습과 저장된 실제 교전 데이터베이스와 비교하여, CO는 불가능할 정도로 축적된 지식 수준에 접근할 수 있다.

전쟁에서 AI를 사용하는 것은 몇 가지 법적, 윤리적 문제를 제기한다. 한 가지 우려되는 것은 인간의 판단을 배제하는 자동화된 무기체계가 국제인도주의법을 위반할 수 있고, 우리의

기본적 생명권과 인간의 존엄성을 위협할 수 있다는 것이다. AI는 또한 전쟁에 돌입하는 문턱을 낮추어 세계 안정에 영향을 미칠 수 있다.

국제 인도주의 법은 어떤 공격도 전투원과 비전투원을 구별해야 하고 비례해야 하며 민간인이나 민가를 대상으로 해서는 안 된다고 규정하고 있다. 또한, 어떠한 공격도 불필요하게 전투원들의 고통을 악화시켜서는 안 된다. AI는 인간의 판단 없이는 이러한 원칙을 이행할 수 없을 수 있다. 특히 많은 연구진은 살상력을 이용해 독자적으로 목표물을 검색하고 '참여'할 수 있는 자율군사 로봇의 일종인 '살상무기체계(LAWS)'가 민간인과 전투원을 구분하지 못하고 민간인 피해를 감안할 때 공격력이 비례하는지 판단할 수 있는지의 여부에 대해 국제인도법이 정한 기준을 충족하지 못할 수 있다는 점을 우려하고 있다.

아모로소(Amoroso)와 탐부리니(Tamburrini)는 'LAWS는 적어도 유능하고 양심적인 인간 군인만큼 차별과 비례의 원칙을 존중할 수 있어야 한다'고 주장한다(2016, p.6). 하지만 림(Lim)은 이러한 요건을 충족하지 못하는 LAWS는 배치해서는 안 되지만 언젠가는 차별성과 비례성 요건을 충족할 만큼 정교해질 것이라고 지적한다(2019). 한편, 아사로(Asaro)는 LAWS가 얼마나 좋아질지 왈가왈부하는 것은 중요하지 않다고 주장한다(2012). 인간만이 무력을 행사해야 한다는 것은 도덕적 요구 사항이며, 기계에 생사 결정을 위임하는 것은 도덕적으로 잘못된 것이다.

인간을 죽이겠다는 결정을 기계에 위임하는 것은 기본적인 인간의 존엄성을 침해하는 것이라고 주장하는 사람들도 있는데, 로봇이 감정을 느끼지 못하기 때문이며, 희생에 대한 개념과 생명을 빼앗는다는 것이 무엇을 의미하는지 알 수 없기 때문이다. 림(Lim) 외 연구진의 설명처럼 '생명체가 아니고 도덕성이나 사망에 대한 개념이 없는 기계는 인간을 상대로 폭력을 쓰는 것의 의미를 헤아릴 수 없고 결정의 중대성에 정의를 내릴 수 없다'는 것이다(2019).

로봇들은 또한 '잘못된' 사람을 죽이는 것이 무엇을 의미하는지 전혀 알지 못한다. 인간만이 인간을 죽이는 것에 수반되는 분노와 괴로움을 느낄 수 있기 때문에 비로소 인간을 상대로 한 희생과 무력 사용을 이해할 수 있다. 또한 '살인에 대한 결정의 중대성'을 이해할 수 있다(Johnson and Axinn, 2013, p 136).

하지만, 다른 사람들은 기계에 의해 죽는 것이 유도 미사일 공격으로 죽는 것보다 주관적으로 더 나쁘거나 덜 존엄한 경험이 될 특별한 이유는 없다고 주장한다. 중요한 것은 피해자가 살해되는 과정에서 굴욕감을 느끼느냐이다. 잠재적 폭격으로 위협을 받는 피해자들은 폭탄이 사람이 떨어뜨리든 로봇에 의해 떨어졌든 상관하지 않을 것이다(Lim et al, 2019). 게다가, 모든 인간이 희생이나 위험에 빠졌다는 감정을 개념화할 수 있는 정서적 능력을 가지고 있는 것은 아니다. 전투의 열기 속에서, 군인들은 희생의 개념에 대해 생각할 시간을 거의 갖지 못하거나, 그들이 무력을 가할 때마다 의식적인 결정을 내릴 수 있는 감정이 생기지 않는다.

또한, 시스템의 사령관, 프로그래머 또는 운영자 중 자율적인 시스템의 조치에 대해 누가 책임을 져야 하는가? 슈밋(Schmit)은 AI를 프로그래밍한 개인과 지휘관 또는 감독관(자율무기 시스템이 전범을 위해 프로그래밍되어 사용됐다는 사실을 알았거나 알았어야 했으며, 이를 막기 위해 아무 것도 하지 않았다고 가정한 경우)이 전범행위의 책임을 져야 한다고 주장한다(2013).

4. AI 표준 및 규정

인공지능과 로봇공학의 윤리적, 법적, 사회적 영향이 더욱 인정받음에 따라 소수의 새로운 세대의 윤리적 표준이 등장하고 있다. 표준이 명시적 또는 암시적 윤리적 우려를 명확히 설명하든, 모든 표준은 어떤 종류의 윤리적 원칙을 구현한다(Winfield, 2019a). 존재하는 표준들은 여전히 개발 중에 있고 그것들에 대한 제한된 공개적으로 이용 가능한 정보들이 있다.

아마도 로봇 공학에서 가장 초기의 명시적 윤리 표준은 로봇과 로봇 시스템의 윤리적 설계와 적용에 대한 BS 8611 가이드일 것이다(British Standard BS 8611, 2016). BS8611은 실행 코드가 아니라 설계자가 잠재적인 윤리적 위해를 식별하고 로봇이나 AI의 윤리적 위험 평가를 수행하며 식별된 윤리적 위험을 완화할 수 있는 방법에 대한 지침이다. 이 지침은 사회, 응용, 상업 및 금융 및 환경이라는 네 가지 범주로 분류된 20개의 고유한 윤리적 위험에 기초한다.

각 위험의 영향을 완화하기 위한 조치와 그러한 조치를 검증할 수 있는 방법에 대한 권고가 있다. 사회적 위험에는 예를 들어 신뢰의 상실, 사기, 사생활 침해, 기밀성 침해, 중독, 고용 상실이 포함된다. 윤리적 위험 평가는 또한 예측 가능한 오남용, 스트레스와 두려움(및 그 최소화), 제어 실패(및 관련 심리적 영향), 재구성 및 책임에 연결된 변경 사항, 특정 로봇 애플리케이션과 관련된 위해성을 고려해야 한다. 학습할 수 있는 로봇과 로봇 발전의 시사점에 특히 관심이 주목되고 있으며, 윤리 표준에서는 로봇 사용과 관련된 윤리적 위험이 인간이 수행할 때와 동일한 활동의 위험을 초과해서는 안 된다고 주장한다.

영국 표준 BS 8611은 물리적 위험이 윤리적 위험을 암시한다고 가정하고, 윤리적 해를 '심리적, 사회적, 환경적 행복'에 영향을 미치는 것으로 정의한다. 또한 물리적 및 정서적 위해성이 사용자에게 예상되는 편익과 균형을 이루어야 한다는 것을 인지하고 있다.

이 표준은 공공부문과 이해관계자를 로봇 개발에 참여시켜야 할 필요성을 강조하고 다음과 같은 주요 설계 고려사항 목록을 제공한다.

- 로봇은 주로 인간을 죽이도록 설계되어서는 안 된다.
- 인간은 책임 있는 대리인으로 남는다.
- 로봇의 책임자를 파악할 수 있어야 한다.
- 로봇은 안전하고 목적에 부하해야 한다.
- 로봇은 사기를 위해 설계되어서는 안 된다.
- 예방원칙에 따라야 한다.
- 설계에 개인정보 보호를 내장하여야 한다.
- 사용자는 로봇을 차별하거나 로봇을 사용하도록 강요당해서는 안 된다.

로봇 공학자들, 특히 연구를 수행하는 사람들을 위해 특별한 지침이 제공된다. 여기에는 대중을 참여시키고, 대중의 관심을 고려하고, 다른 분야의 전문가들과 협력하고, 잘못된 정보를 수정하고 명확한 지침을 제공해야 할 필요성이 포함된다. 로봇의 윤리적 사용을 보장하기 위한 구체적인 방법에는 사용자 검증(예상대로 로봇이 작동할 수 있도록 보장), 소프트웨어 검증(예상대로 소프트웨어가 작동되도록 보장), 윤리적 평가에 다른 전문가의 참여, 예상 결과에 대한 경제적, 사회적 평가, 법적 영향 평가, 준수 테스트 등이 포함된다. 관련 표준에 어긋나다 해당하는 경우 로봇의 설계와 운용에 있어 다른 지침과 윤리 규정을 고려해야 한다(예: 특

정 상황에 관련된 의료 또는 법률 코드). 이 표준은 또한 로봇의 군사적 적용에 있어 인간의 책임과 의무를 없애지 않는 경우를 만든다.

IEEE 표준 협회는 또한 자율 및 지능형 시스템의 윤리에 관한 글로벌 협회를 통해 윤리 표준을 발표했다. IEEE협회는 '인간 웰빙'을 중심 지침으로 배치하면서 로봇과 AI를 단순한 경제성장을 위한 도구가 아닌 인간 조건 개선을 위한 기술로 재배치할 방법을 명시적으로 모색한다(Winfield, 2019a). 그들의 목적은 AI와 로봇의 이해당사자들이 '윤리적인 고려에 우선순위를 두어 인류를 위해 이 기술들이 진보하도록' 교육하고, 훈련하며, 권한을 부여하는 것이다.

현재 14개의 IEEE 표준 연구 그룹이 인공지능에 영향을 미치는 이른바 '인간' 표준의 초안을 작성하고 있다(표 4.1).

표 2 : AI를 위한 IEEE '인간 표준'

표준		목표
P7000	시스템 설계에서 발생하는 윤리적 문제와 대처에 관하는 모델 프로세스	자율 및 지능형 시스템의 윤리적인 설계를 위한 프로세스 수립
P7001	자율 시스템의 투명성	다양한 이해 관계자에게 자율 시스템의 투명성을 보장. • 사용자: 신뢰를 주기 위해 시스템이 무엇을 하고 왜 하는지 사용자가 이해할 수 있도록 보장 • 검증 및 인증: 시스템이 정밀 검사를 받도록 보장 • 사고: 사고 조사관이 조사에 착수할 수 있도록 보장 • 변호사 및 전문가 증인: 사고 발생 후 증인들이 증거를 제시할 수 있도록 보장 • 파괴적 기술(예: 운전자 없는 자동차): 대중이 기술을 평가할 수 있도록 보장(해당되는 경우, 신뢰도 구축).
P7002	데이터 개인 정보 보호 프로세스	소프트웨어 공학 프로세스에서 개인 데이터의 윤리적 사용에 대한 표준을 제정. 프라이버시 통제 조치의 필요성과 효과를 식별하는 데 사용될 수 있는 프라이버시 영향 평가(PIA)를 개발하고 설명하고, 개인 정보를 사용하는 소프트웨어를 개발하는 사람들에게 체크리스트를 제공하게 될 것
P7003	알고리즘 편향 고려사항	알고리즘 개발자가 제품의 편중 위험을 제거하거나 최소화하고자 했던 방법을 명시적으로 작성할 수 있도록 도움. 이는 지나치게 주관적인 정보의 사용을 다루고 개발자가 보호 특성(예: 인종, 성별)에 관한 법률을 준수하는지 확인하는 데 도움이 될 것이다.

		- 데이터셋 선택을 위한 벤치마킹 프로세스 - 구축되고 검증된 알고리즘의 범위 전달에 관한 지침(예상하지 않은 사용의 의도하지 않은 결과로부터 보호) - 사용자가 시스템 출력을 잘못 해석하지 않도록 하기 위한 전략
P7004	아동 및 학생 데이터 거버넌스 표준	특히 교육기관을 대상으로 한 이 지침은 아동/학생 데이터의 접근, 수집, 저장, 사용, 공유 및 파기에 대한 지침을 제공
P7005	투명한 고용인 데이터 거버넌스 표준	P7004와 비슷하지만 고용인을 대상으로 함
P7006	개인 데이터 인공지능(AI) 에이전트 표준	개인화된 AI에 대한 액세스를 생성하고 부여하는 데 필요한 기술적 요소를 설명함. 개인이 기계 판독 가능한 수준에서 개인정보를 안전하게 정리하고 공유할 수 있고, 개인화된 AI가 기계 대 기계 의사 결정의 대리 역할을 할 수 있게 됨.
P7007	윤리적 로봇공학과 자동화 시스템의 존재론적 표준	이 표준은 제품 수명 주기 전체에 걸쳐 사용자 웰빙을 고려하기 위해 엔지니어링과 철학을 결합함. 이익을 극대화하고 부정적인 영향을 최소화하는 방법을 찾고, 다양한 공동체들 간에 의사소통을 명확히 할 수 있는 방법도 고려하고자 함.
P7008	로보틱, 인텔리전트, 자동화 시스템의 확산을 윤리적으로 주도할 수 있는 표준	'넛지 이론'을 바탕으로, 이 표준은 로봇이나 자율 시스템이 수행할 수 있는 현재 또는 잠재적 넛지를 설명하고자 함. 넛지는 다양한 이유로 사용될 수 있지만, 수신자에게 정서적으로 영향을 주고 행동을 변화시키며 조작적일 수 있음을 인식하고 넛지를 이용한 AI의 윤리적 설계를 위한 방법론을 정교하게 다듬으려고 함.
P7009	자동화 및 반자동화 시스템의 비상안전 설계 표준	강력하고 투명하며 책임 있는 페일세이프 메커니즘의 개발과 구현을 위한 효과적인 방법론을 만듦. 시스템의 고장 능력을 안전하게 측정하고 테스트하는 방법을 설명함.
P7010	윤리적 AI 및 자동화 시스템의 복리성 측정 표준	자율 시스템에 의해 영향을 받을 수 있는 웰빙 요소를 평가하는 데 사용되는 측정 기준과 인간의 웰빙이 능동적으로 개선될 수 있는 방법에 대한 기준을 설정함.
P7011	뉴스 출처의 신뢰성을 측정 및 확인하는 절차를 위한 표준	뉴스 정보에 초점을 맞춘 이 표준은 뉴스 기사의 사실적 정확성을 평가하기 위한 프로세스를 표준화하기 위해 발표되었고, '신뢰성' 점수를 내기 위해 사용될 것임. 이 표준은 확인되지 않은 '가짜' 뉴스의 부정적인 영향을 해결하고자 하며, 뉴스 제공자에 대한 신뢰를 회복하기 위해 고안됨.

P7012	기계 판독 가능한 개인 정보보호 조건을 위한 표준	개인 정보 보호 관련 용어가 어떻게 제시되는지, 그리고 어떻게 기계에서 읽고 받아들일 수 있는지를 확립하기 위함.
P7013	자동 안면 분석 기술을 위한 포함 및 적용 표준	안면 인식에 사용되는 데이터, 다양성 요건, 얼굴 인식이 사용되어서는 안 되는 애플리케이션 및 상황의 벤치마킹에 대한 지침 제공