ISSUE REPORT I 2021.03.25. IS-114

유럽(EU)의 인공지능 윤리 정책 현황과 시사점 : 원칙에서 실천으로

An European approach to ethics on artificial intelligence: from principles to practices

유재흥, 추형석, 강송희

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

이 보고서는 「과학기술정보통신부 정보통신진흥기금」에서 지원받아 제작한 것으로
과학기술정보통신부의 공식의견과 다를 수 있습니다.
이 보고서의 내용은 연구진의 개인 견해이며, 본 보고서와 관련한 의문 사항 또는 수정·보완할 필요가 있는 경우에는 아래 연락처로 연락해 주시기 바랍니다.

소프트웨어정책연구소 AI정책연구팀
유재흥 선임연구원 jayoo@spri.kr

CONTENT

요 약 문

유럽(EU)은 2018년 4월 유럽AI전략을 수립한 이후 이의 실현을 위한 일련의 정책을 추진 중이다. 2019년 4월 발간한 『신뢰할 만한 AI윤리 가이드라인』과 2020년 7월 공개한 AI 윤리 『평가 목록(assessment list)』은 이러한 노력의 결과다. 이 작업은 52인으로 구성된 AI고위전문가그룹(AI-HLEG)과 유럽AI연합회(European AI Alliance)를 통해 이루어졌다. AI고위전문가 그룹이 작성한 보고서 초안은 4,000여명의 이해관계자가 참여하는 유럽AI연합회를 통해 검토되었고 다양한 제안들이 최종 보고서에 반영되었다. 이 과정은 AI 윤리의 중요성과 핵심 AI 윤리 원칙, 실무에 있어서 구체적 고려사항을 산업계와 시민사회와 공유함으로써 내용과 절차의 투명성을 확보했다는 의미를 갖는다.

EU의 AI 윤리 가이드라인은 인간의 기본권 존중을 AI 개발의 가장 기본적 원칙으로 삼고, 신뢰할 만한 AI의 속성으로 적법성(lawful), 윤리성(ethical), 견고성(robust)을 제안한다. 나아가 신뢰할 만한 AI의 7대 요구사항을 인간행위자와 감독, 기술적 견고성과 안전, 프라이버시와 데이터 거버넌스, 투명성, 다양성·차별금지·공정성, 사회·환경적 복지, 책임성 측면에서 정리했다. 이러한 요구 사항이 AI 개발 수명 주기에서 잘 충족이 되는지를 140여 가지 세부 평가 목록을 활용해 실무에서 점검할 수 있도록 하였다. 특히 실천적 AI 윤리를 위한 평가 목록은 기존의 원칙 선언 위주의 보고서와는 차별되는 EU AI고위전문가그룹의 기여라 평가된다.

우리나라는 2019년 12월 『인공지능 국가전략』을 발표하고, 일년 만인 2020년 12월 AI 윤리 기준을 마련하였다. 지난 2021년 1월 발생한 AI챗봇 서비스 '이루다' 사건은 AI 윤리 이슈의 중요성과 파급력을 부각시키는 계기가 되었다. 이제 AI 윤리는 원칙 선언에서 실천으로 옮기는 과정에 있다. 정부가 계획 중인 AI 윤리 체크리스트 마련이 대안이 될 수 있다. 다만 그 과정에서 다양한 이해관계자의 의견 수렴과 국제 사회와의 공조 및 관련 단체와 표준에 대한 지속적 동향 파악을 통해 윤리적인 AI 기술과 산업 혁신 촉진 사이의 균형 있는 정책적 접근이 요구된다.

Executive Summary

Since the European AI strategy was published in April 2018, Europe has been pursuing a series of policies to implement it. The『Ethics Guidelines for Trustworthy AI』released in April 2019 and the『Assessment List』for self-inspection released in July 2020 are the results of such efforts. EU's Ethics Guidelines are proposed as the most fundamental principles of AI development, respecting human rights, and as attributes of trustworthy AI, lawfulness, ethicalness, and robustness. More specifically, the seven requirements of trustworthy AI are organized into human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination, fairness, social and environmental well-being, and accountability. The list of assessment consisted of 140 specific questions that can confirm whether or not the basic rights of humans are violated and whether the seven ethical requirements have been met.

Korea established a『National strategy for Artificial Intelligence』in December 2019 and announced AI ethics standards in December 2020, a year later. The AI chatbot service "Iruda" incident, which took place in January 2021, highlighted AI ethics issues. It is now necessary to come up with policies that can be implemented from the AI principle declaration. An alternative is to prepare an AI ethics checklist that the government is planning. However, in the process, a balanced policy approach between protecting personal information and promoting industrial innovation is required by collecting opinions from various stakeholders, cooperating with the international community, and grasping regional trends in related organizations and standards.

I. 서론

1. 연구 배경

□ 인공지능(AI)에 대한 관심이 높아지면서 AI 제품·서비스의 개발 및 상용화 과정에서 윤리적 문제가 사회적 이슈로 부상함

○ AI의 데이터 학습 과정에서의 편향성, 결과 도출 과정의 불투명성, 문제 발생 시 책임소재의 불명확성으로 인한 잠재적 위험성*이 지속적으로 제기되어 왔음

* (6대 위험) 편견과 차별, 개인의 자율성·의지·권리 보장 부정, 불투명·불명확·부정당한 결과, 사생활 침해, 사회적 관계의 고립과 붕괴, 낮은 수준의 결과 (KPC4IR,'19)

○ 실제로, 구글, 마이크로소프트(MS), 아마존 등 선도 IT기업들의 AI 서비스 개발 및 상용화 과정에서 기계 학습의 한계로 인한 윤리 이슈*가 발생

* (아마존) 채용 AI의 남성 선호 이슈로 도입 취소('15), (구글) 포토 서비스가 흑인을 고릴라로 태깅('15), (MS) 챗봇 테이(TAY)는 욕설·극단적 발언 학습('16)

○ 올 초 국내 한 스타트업의 인공지능 챗봇 서비스의 윤리 문제*가 불거지면서 사회적인 논란을 촉발함

* AI 스타트업 스캐터랩에서 출시한 인공지능 챗봇 서비스(이루다)가 성희롱·혐오발언으로 사회적 논란을 일으키며 현재 서비스가 중단된 상태('21.1)

□ 선도 IT기업들은 물론 국제 사회와 해외 주요 국가에서는 AI 윤리 원칙 및 가이드라인 마련 등 관련 대응을 해오고 있음

○ 기업들은 자체적으로 AI 윤리 원칙과 가이드라인 마련하여 시행중*

* (구글) AI 원칙 제정('18.6), (MS) 마이크로소프트 책임있는 AI 원칙 제정 ('16) (카카오) AI 윤리 헌장 제정('18.1) (네이버) AI 윤리 준칙 발표 ('21.2)

○ 미국, 일본, 영국 등 해외 주요국 정부 및 국제 사회에서도 AI 윤리 원칙 및 가이드라인을 마련하고 공개하고 있음

[표1-1] 기업 및 정부 기관들의 AI윤리 대응 동향

구분	주체	AI윤리 관련 정책 대응 동향
기업	카카오	• 카카오 알고리즘 윤리헌장 제정('18.1) • 전 직원 대상 AI윤리 교육 추진('21.2)
	삼성전자	• 국내 기업 최초 Partnership on AI 가입('18.11) • AI 윤리 기준 제정 준비 중
	네이버	• 서울대와 공동으로 AI윤리 준칙 발표('21.2)
	구글	• 구글 AI 원칙 발표('18.6)
	마이크로소프트	• 마이크로소프트 책임 있는 AI 원칙 발표('20)
정부	미국	• 백악관, AI 이니셔티브('19.2)
	일본	• AI 개발 원칙('17)[1], AI 사회 원칙('19)
	영국	• 데이터 윤리 프레임워크('18.8) • 알란튜링연구소, 인공지능 윤리와 안전성의 이해 가이드라인('19) • 정보위원회(ICO), 인공지능 및 데이터 보호 가이드라인('20.7)
	한국	• 인공지능 윤리 기준('20.12) 발표
국제사회	OECD	• OECD AI 원칙('19.5), G20에서 OECD AI 원칙 수용('19.6)
	EU	• 신뢰할 만한 인공지능 윤리 가이드라인('19.4) 및 평가 목록 발표('20.7)
	UNESCO	• 인공지능 윤리에 대한 권고사항 초안('19.5)[2]

2. 보고서 목적

□ 우리나라는 인공지능 국가전략('19.12) 수립 이후 1년 만에 인공지능 윤리기준(안)을 마련('20.12)하고 세부 계획들을 추진 중

○ 지난 1년간 정부 차원의 AI 윤리연구반을 구성하여 초안을 마련하고 전문가 의견 수렴과 공청회 등을 거쳐* 최종안을 확정 발표하였음('20.12)

* 연구반 구성 및 운영 ('20.4~8) → 각계 의견 수렴 ('20.9~12) → 최종안 공개 의견 수렴 (공청회(12.7), 전문가 및 시민 의견 접수(11.25~12.5)) → 발표(12.23)

1) NIA('17.9), AI연구개발과 활용촉진을 위한 『AI 개발 가이드라인(안)』

2) UNESCO('19.5), Elaboration of a recommendation on the ethics of artificial intelligence

○ 향후 AI 윤리기준을 현장에서 실천할 수 있도록 주체별 체크리스트, 교육 프로그램 및 필요시 윤리 보완 및 세부 기준 및 입법 지원 계획

- AI 개발에서 활용까지 전 과정에 참여하는 주체별 AI 윤리 점검을 위한 체크리스트를 개발·보급하고 AI 윤리 커리큘럼도 마련 예정

□ 이러한 상황에서, 최근 발표한 유럽연합(EU)의 『신뢰할 만한 AI 윤리 가이드라인』과 자율 점검을 위한 『평가 목록』의 개발 과정과 내용은 참조할 필요가 있음

○ (전략 정합성) EU의 AI 윤리 가이드라인과 평가 목록의 개발은 '18년 발표한 유럽의 AI 전략을 실현하기 위한 기본 정책의 하나로 추진됨

- EU의 법규범, 디지털 단일 시장 구현이라는 공동의 목표 하에 AI 산업 활성화를 추진하되 신뢰 기반으로서 AI 규범 체계를 우선적으로 정립하는데 목적

○ (추진 체계) 전문가 그룹(AI-HLEG)과 의견 수렴 플랫폼(유럽AI연합회)을 조직해 보고서 작성과 광범위한 의견 수렴을 병행

- 전문가 그룹이 작성한 보고서의 초안은 4,000여명의 시민, 전문가, 기관 등이 참여하고 있는 유럽AI연합회가 검토하고 의견을 제시

- 각계 이해관계자 의견이 반영된 수정된 보고서는 유럽위원회(European Commission)에 보고되고 입법 영향 평가를 거쳐 입법 제안서로 준비됨

○ (과정의 투명성) 보고서의 작성 과정과 수렴된 의견, 최종 산출물은 온라인을 통해 모두 공개되고 있어 추진 과정과 내용을 투명하게 알 수 있게 함

- 가이드라인 초안 및 초안에 대한 500여건 검토 의견, 평가 목록의 초안과 평가 목록의 파일럿 테스트 방법 등이 온라인으로 공개되어 있음

- 140여 가지의 항목으로 구성된 평가 목록(assessment list)도 온라인 툴킷(tool kit)으로 공개해 일반인들의 접근성과 활용성을 높이고 있음

□ 이 보고서에서는 EU의 AI 윤리 가이드라인 및 평가 목록의 개발 과정과 내용을 보다 구체적으로 살펴보고 향후 국내 AI 윤리 원칙 및 정책 추진의 시사점을 얻고자 함

II. 유럽의 AI 윤리 정책 추진 현황

1. 추진 현황

□ 유럽은 '18년부터 『유럽 인공지능 전략』('18.4), 『신뢰할 만한 인공지능 윤리 가이드라인』('19.4), 『인공지능 백서』('20.2)등의 AI 정책 보고서를 발간함

○ 지난 '18년 4월 유럽의 AI 전략 보고서 발간 이후 이를 실현하기 위해 후속 정책 보고서, 회의, 공개 의견 수렴, 법 제도 논의가 이뤄지고 있음

○ 그간의 AI 관련 다양한 논의의 결과를 정리해 '21년 1분기에 AI 관련 새로운 입법 제안서를 완성할 예정임

[표2-1] 유럽의 인공지능 전략 주요 추진 경과

시기	추진 내용
2021.1Q	• 입법 제안서 제출 (Legislative proposal on AI) 계획
2021.1Q	• 유럽 AI 협력 계획 업데이트 (Updated Coordinated Plan on AI) 계획
2020.12	• 영향평가(Impact Assessment) 보고서
2020.07	• 초기영향평가(Inception Impact Assessment) 보고서
	• '인공지능 백서(White Paper on Artificial Intelligence)' 공개의견수렴
	• AI-HLEG, 신뢰할만한 AI를 위한 최종 평가(assessment list) 목록 공개
	• AI-HLEG, 신뢰할만한 AI를 위한 분야별 권고안 발간
2020.02	• '인공지능 백서: 수월성(excellence)과 신뢰(trust)를 위한 유럽의 접근' 발간
2019.12	• 신뢰할만한 AI를 위한 평가 목록 파일럿 테스트
2019.06	• 제1회 유럽AI연합회 의회 (1st European AI Alliance Assembly) 개최 • AI-HLEG, 정책 및 투자 권고안 (Policy and Investment recommendations)
2019.04	• 신뢰할 만한 AI 윤리 가이드라인 발간 (Ethics Guidelines for Trustworthy AI)
2018.12	• 유럽 AI 협력 계획 발표 (Communication on " AI Made in Europe") • '신뢰할 만한 AI 윤리 가이드라인' 이해관계자 정책 제언 수렴
2018.06	• 유럽AI위원회 (European AI Alliance) 출범 • AI고위전문가그룹 조직(High Level Expert Group on AI, AI-HLEG)
2018.04	• 유럽 AI 전략 발표 (Communication: Artificial Intelligence for Europe) • 유럽 AI 협력 선언(Declaration of cooperation on Artificial intelligence)

2. 추진 체계

□ 유럽위원회(European Commission)가 발족한 AI-HLEG*가 가이드라인 작성을 담당했으며, 동시에 조직한 유럽 AI 연합회(European AI Alliance)를 통해 보고서 검토 및 수정·보완을 위한 이해관계자의 의견을 수렴함

* AI-HLEG(High Level Expert Group on AI)은 52인의 산학연 AI 전문가로 구성된 정책 자문 기구로 '18년 설립되어 유럽의 『신뢰할 만한 AI 윤리 가이드라인』 및 신뢰성 있는 AI 개발을 위한 정책 보고서를 작성함[3)]

○ (AI-HLEG) '18년 4월 『유럽 인공지능 전략』 발표 이후 유럽위원회에 의해 구성되어, AI 윤리 가이드라인 및 점검평가 목록을 작성한 후 '20년 7월 활동을 종료함

- '19.4월: AI 윤리 가이드라인 수정본 발표
- '19.6월: 제1회 유럽 AI 연합회 의회[4)]에서 AI 윤리 가이드라인을 발표한 후 유럽위원회는 AI-HLEG의 활동을 1년 연장 함
- '20.7월: 평가 목록의 파일럿 테스트 및 의견 수렴을 통해 AI 윤리 자율 점검 평가 목록 최종 완성 후 활동 종료

○ (유럽AI연합회) AI-HLEG는 보고서 작성을 위해 유럽 AI 연합회를 다양한 정책 의제와 의견을 투명하게 수렴하는 창구로 활용

- 유럽 AI 연합회는 4천여 명의 유럽 시민과 다양한 이해관계자로 구성된 포럼으로 AI-HLEG와 함께 조직되어 다양한 정책 의견 수렴을 지원하였음

※ 지난해 10월 개최된 제2회 유럽AI연합회 의회는 온라인으로 진행되었고 유럽의 AI 생태계 경쟁력 강화와 신뢰할 만한 AI 개발을 위한 발표와 토론이 이뤄짐

3) EU AI High Level Exppert Group on AI,
(url : https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence)

4) First EU AI Alliance Assembly,
(url : https://ec.europa.eu/digital-single-market/en/news/first-european-ai-alliance-assembly)

III. 유럽의 AI 윤리 정책의 주요 내용

1. 『신뢰할 만한 AI 윤리 가이드라인』('19.4)

□ (배경) 유럽위원회는 '18년 4월 25일[5]과 12월 7일자[6] 두 차례 연락문을 통해 "유럽에서 만들어진 윤리적이고 안전한 첨단 인공지능"을 지지하는 AI에 대한 비전을 선포함

○ 위원회 비전의 세 가지 중심축은 △ AI 수용 확대를 위한 공공 및 민간 투자 증대, △ 사회경제적 변화 대비, △ 유럽의 가치 강화를 위한 적절한 윤리적·법적 체계 확보임

□ (목적) 가이드라인은 적법하고, 윤리적이며, 견실한 AI 시스템의 설계, 개발, 이용을 위한 원칙, 요구사항 및 자율 점검을 위한 평가 목록 개발을 목적으로 함

○ EU 조약, EU 헌장 및 국제 인권법에 규정된 기본권을 바탕으로 AI의 개발, 배포, 이용에 의해 제기되는 윤리적 문제에 초점을 둠

□ (추진 경과) '18년 12월 초안 완료 후 공개 의견을 반영해 '19년 4월 수정본이 발표되었고 보고서에 포함된 AI 윤리 평가 목록은 '20년 7월 최종 공개 됨

○ ('18.12) 신뢰할 수 있는 AI 윤리 가이드라인 초안을 공개함

○ ('18.12~'19.2) 유럽AI연합회가 공개 의견 수렴 과정을 통해 EU 소속국과 그 외 국가로부터 총 506개의 제안(contributions)* 취합[7]

* 초안 문서의 각 챕터(Chapter)별 의견 및 제안을 받은 결과 총 506건의 의견 중 337개는 공개 의견**, 97개는 익명 의견, 72개는 비공개 의견

** 337개의 공개 의견 중 244개는 조직, 소속, 회사, 직책을 명시, 49개는 일반적 소속 명시(교수, 연구자, 대학원 생 등), 44개는 개인 의견

5) European AI Strategy (Communication: Artificial Intelligence for Europe - Press Release)

6) Coordinated Plan on AI (Communication on " AI Made in Europe" - Press Release)

7) 주요 의견 요약은 별첨 자료 참고

○ (‘19.1~‘19.3) AI-HLEG 보고서 초안 수정 회의*를 개최함

* 벨기에 브뤼셀에서 총 5차례(‘19.1.22~23, 2.25, 3.18~19)에 걸쳐 검토 및 수정

○ (‘19.3) AI-HLEG가 채택한 수정보고서를 유럽위원회에 전달함

○ (‘19.4) 수정 보고서(Ethics guidelines for Trustworthy AI) 공개함

□ (주요내용) AI 또는 AI 시스템에 대한 개념을 정의하고 신뢰할 수 있는 AI의 필수 3요소, 7대 요구사항, 구현을 위한 기술적·비기술적 방안을 제안함

○ (AI관련 정의) 보고서에 사용된 AI 또는 AI 시스템이라는 용어의 일관적 사용을 위해 AI-HLEG에서 별도의 ‘AI 정의(definition)’ 문서를 작성함[8)]

- (AI정의) 인공지능은 특정한 목적 달성을 위해 환경 분석을 수행하고 어느 정도의 자율성(autonomy)을 기반으로 지능적 행동을 수행하는 시스템
- (AI시스템) AI 기반 시스템은 음성·이미지 인식 소프트웨어와 같은 순수 소프트웨어이거나 자율주행차, 로봇과 같이 AI가 탑재된 하드웨어일 수 있음

○ (3대 요소) 신뢰할 수 있는 AI의 필수 3 요소로 적법성, 윤리성, 견고성 제안

- (적법성, Lawful) 법적 구속력이 있는 모든 법률과 규정 준수
- (윤리성, Ethical) 구속력이 없더라도 윤리적 원칙과 가치에 부합해야 함
- (견고성, Robust) 선한 의도로 개발된 AI라도 기술적·사회적 관점에서 의도치 않은 피해를 유발하지 않을 것이라는 기대에 부합해야 함

○ (7대 요구사항) 인간행위자와 감독, 기술적 견고성 및 안전, 프라이버시와 데이터 거버넌스, 투명성, 다양성·차별금지·공정성, 사회·환경적 복지, 책임성

① (인간 행위자와 감독) AI 시스템은 인간의 자율성과 의사결정을 보조하는 역할을 수행하고, AI 시스템에 대한 인간의 감독권을 허용해야 함

② (기술적 견고성과 안전성) AI 시스템에 대한 외부 공격 대응, 오류 발생시 체계적 대처, AI 기능의 정확성과 재현성 확보, 의도치 않은 결과와 작동 오류를 최소화해야 함

8) AI-HLEG(2019.4.8.), A definition of Artificial Intelligence: main capabilities and scientific disciplines

③ (프라이버시와 데이터 거버넌스) AI 시스템 전주기의 개인정보 및 데이터 보호, 학습용 데이터의 품질 확보와 문서화, 개인정보 처리시 데이터 접근 규정 마련

④ (투명성) AI 시스템의 데이터 수집·가공 및 학습 과정의 문서화를 통한 추적 가능성 제고, AI 시스템에 대한 설명요구권 대응, AI 시스템의 인지·역량·한계를 사용자에게 제공

⑤ (다양성, 차별 금지, 공정성) 불공정한 편향성 방지 및 다양성 확보를 위한 노력, 연령·성별·국적·장애 요인에 무관한 보편적 설계, AI 시스템과 연관된 이해관계자와의 참여 독려 및 협의

⑥ (사회적, 환경적 웰빙) 환경친화적 AI 시스템 개발과정 수립, AI 시스템의 사회적 역기능 방지를 위한 모니터링, 정치적 의사결정 및 선거 상황 등에서의 AI 시스템 활용에 신중한 고려

⑦ (책임성) AI 시스템 개발·배포·이용 과정상에 발생할 수 있는 책임을 이행하기 위해 감사, 영향평가, 상충 관계 평가, 구제방안 수립

[표3-1] EU의 신뢰할 만한 AI의 7대 요구사항

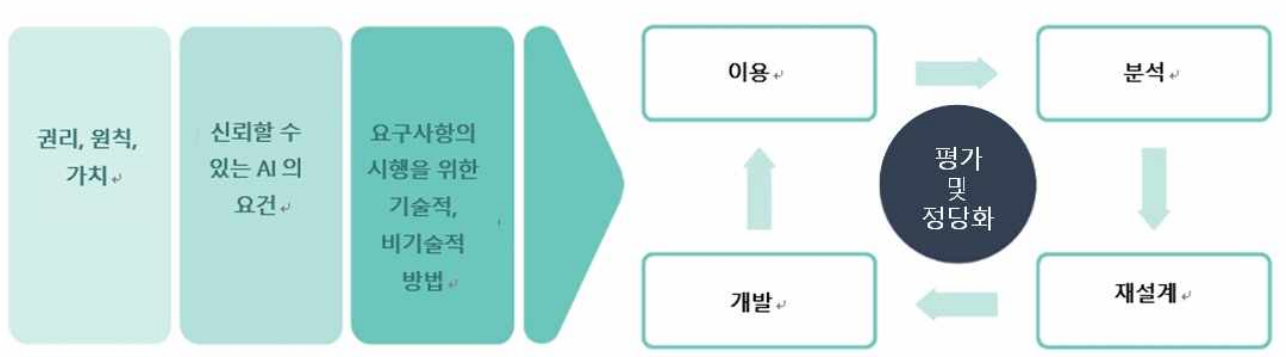
요구사항	내용
인간 행위자와 감독 Human agency and oversight	• AI 시스템은 인간의 자율성과 의사결정을 뒷받침하는 역할로 사용자의 주체성을 뒷받침함으로써 번성하고 평등한 민주주의 사회를 가능하게 하는 역할을 하고, 기본권을 증진하며, 인간의 감독을 허용해야 함 • 감독은 인간의 개입(HITL: human-in-the-loop), 인간의 감시(HOTL: human-on-the-loop), 인간의 명령(HIC: human-in-command)과 같은 거버넌스 체계를 통해 달성 가능 △ HITL은 시스템의 모든 의사결정 주기에 인간이 개입할 수 있는 기능을 의미하며, 많은 경우 이는 가능하지도, 바람직하지도 않음 △ HOTL은 시스템의 설계 주기에 대한 인간의 개입과 시스템 운영에 대한 모니터링 기능을 의미 △HIC는 AI 시스템의 전반적인 활동을 감시하는 기능과 (폭넓은 경제, 사회, 법률, 윤리적 영향을 포함한다) 특정 상황에서 시스템을 언제, 어떻게 사용할지 결정할 수 있는 능력을 의미 • 인간이 AI 시스템에 대해 행사할 수 있는 감독권이 적을수록 보다 광범위한 시험과 엄격한 거버넌스가 요구

기술적 견고성과 안전성 Technical robustness and safety	• 공격에 대한 회복 탄력성과 시스템 에러 및 실패의 대체 계획, 일반적 안전, 정확성, 신뢰성, 재현 가능성 포함 • 기술적 견고성은 AI 시스템의 개발에 있어 위험 예방을 위한 접근방식이 존재하고 의도한 대로 신뢰성 있게 작동하는 한편 의도치 않고 예상하지 못한 위해를 최소화하며 용인 불가능한 위해를 방지하는 방식으로 개발할 것을 요구 • AI 시스템의 의도치 않은 응용(예: 이중 용도 애플리케이션)과 악의적 행위자에 의한 시스템 남용 가능성을 고려해 이를 완화, 방지하기 위한 조치를 취해야 함 • 인간의 신체적, 정신적 무결성도 보장 • AI 시스템은 문제 발생 시 대체 계획을 가능하게 하는 보호조치를 구비해야 함 • 시스템이 생명체 또는 환경에 위해를 가하지 않고 하도록 되어 있는 일을 할 것이라는 점을 반드시 담보해야 함 • 다양한 응용 분야에 걸쳐 AI 시스템의 이용과 관련된 잠재적 위험을 분명히 파악하고 평가하기 위한 절치를 수립
프라이버시와 데이터 거버넌스 Privacy and Data governance	• 개인정보 존중, 데이터의 질과 무결성, 데이터에 대한 접근성 포함 • AI 시스템이 시스템의 수명 주기 전체에 걸쳐 개인정보와 데이터 보호 보장 • 학습에 적합한 충분한 품질의 데이터 세트와 데이터의 무결성을 확보해야 함 • 사용되는 프로세스와 데이터 세트에 대해 반드시 계획, 학습, 시험, 배포와 같은 각 단계에서 시험과 문서화가 이루어져야 하고 이는 자체적으로 개발했으나 다른 곳에서 인수한 AI 시스템에도 적용 • 개인 데이터를 다루는 조직에는 데이터 접근을 규율하는 데이터 프로토콜 확보
투명성 Transparency	• 추적 가능성, 설명 가능성, 의사소통 포함 • 데이터 수집과 라벨링, 사용되는 알고리즘을 포함하여 AI 시스템의 결정을 만들어내는 데이터 세트와 프로세스는 가능한 최선의 표준에 맞게 문서화하여 추적 가능성과 투명성을 향상시켜야 함 • AI 시스템이 사람의 삶에 중대한 영향을 미치는 경우 AI 시스템의 의사결정 과정에 대해 적합한 설명을 요구할 수 있어야 함 • AI 시스템은 사용자에게 자신을 인간으로 표현해서는 안 되며, 인간은 자신이 AI 시스템과 상호작용하고 있다는 정보를 제공받을 권리를 갖음 • AI 관계자 또는 최종 사용자에게 AI 시스템의 역량(정확도 등)과 한계에 대해 알려야 함
다양성, 차별 금지, 공정성 Diversity, non-discrimination and fairness	• 불공정한 편향의 방지, 보편적 설계, 이해관계자의 참여 포함 • 가능한 경우 수집 단계에서 파악할 수 있는 차별적 편향을 제거 • 시스템의 목적과 제약, 요구사항, 의사결정을 투명하고 분명한 방식으로 분석하고 다루는 감시 프로세스를 시행함 • 다양한 배경, 문화, 분야의 인력을 고용함으로써 의견 다양성 확보 • 사용자의 접근성 확보를 위해 특히 기업 대 고객의 영역에서 시스템은 사용자 중심적이어야 하며 연령, 성별, 능력, 장애에 관계 없이 모든 사람들이 AI 제품 또는 서비스를 사용할 수 있도록 보편적 설계 원칙을 고려해야 함 • AI 시스템을 개발하기 위해 수명 주기 전체에 걸쳐 시스템의 영향을 직접적 또는 간접적으로 받을 수 있는 이해관계자들과 협의하는 것이 바람직

사회적, 환경적 웰빙 Societal and environmental well-being	• 지속가능성과 환경 친화성, 사회적 영향, 사회와 민주주의 포함 • (친환경) 학습 과정에서의 자원 이용과 에너지 소비를 철저히 조사하고 보다 덜 해로운 선택을 하는 것 등을 통해 시스템의 개발, 배포 및 이용 과정과 공급 사슬 전체를 환경적 측면에서 평가할 필요 • 우리 삶의 모든 영역에서 (교육, 근로, 양육, 오락 등) 사회적 AI 시스템에 대한 전방위적 노출이 사람의 신체적, 정신적, 사회적 관계에도 영향을 미칠 수 있음을 고려해 시스템 효과를 면밀히 관찰 • AI시스템의 개발, 배포, 이용의 개인적 차원의 영향력뿐만 아니라 제도, 민주주의 등 사회적 관점에 대한 영향력 평가도 이루어져야 함
책임성 Accountability	• 감사 가능성, 부정적 영향의 최소화와 보고, 상충과 구제 포함 • AI 시스템의 개발, 배포, 이용 전후에 AI 시스템과 그 결과물의 책임성을 담보하기 위한 체계를 시행할 것을 요구 • 알고리즘, 데이터 및 설계 과정의 평가를 가능하게 해야 하나 (감사 가능성 확보) 이것이 사업 모델, 지식 재산의 공개를 의미하지는 않음 • AI 시스템의 잠재적인 부정적 영향력을 파악, 평가, 문서화, 최소화 • AI 시스템에 관한 적절한 우려사항을 신고할 때는 내부 고발자, 비정부조직, 노조 또는 기타 단체에 대해 정당한 보호가 제공 • AI 시스템과 관련이 있는 이해관계와 가치를 파악하고 충돌이 발생하는 경우 기본권을 포함한 윤리 원칙에 대한 위험성의 측면에서 상충 관계를 명시적으로 인정하고 평가해야 하며, 윤리적으로 용인 가능한 상충 관계를 파악할 수 없는 상황에서는 그러한 형태로 AI 시스템의 개발, 배포, 이용을 진행시켜서는 안 됨 • 부당한 악영향이 발생하는 경우 충분한 이용자 구제를 가능케 접근 체계 제공

○ (실현 방안) 신뢰할 수 있는 AI 실현을 위한 기술적·비기술적 방안을 제안하고 있으며 동적으로 변화하는 AI 시스템의 특성상 시스템의 전 주기에 걸쳐 지속적으로 시행, 평가, 정당화 과정을 진행하며 보완해 나갈 것을 권고

[그림3-1] 시스템 수명 주기에 걸쳐 신뢰할 수 있는 AI의 실현 과정



- (기술적 방법) 시스템 제한 사항 기반의 아키텍처, 규범 기반의 설계, 설명 가능한 AI 기술 개발, 전 수명 주기에 걸친 시험과 검증 활용

기술적 방법	내용
아키텍처	• 시스템이 항상 따라야 하는 일련의 “화이트리스트(white list)” 규칙 (행동 또는 상태) 또는 시스템이 절대 위반해서는 안 되는 행동 또는 상태에 대한 “블랙리스트(black list)” 제한사항 등을 작성해 시스템 아키텍처 차원에서 윤리적 요구사항 준수
설계 (X-by-Design)	• 설계에 의한 개인정보 보호, 설계에 의한 보안 등과 같이 다양한 “설계에 의한 X” 개념을 적용, 같이 신뢰를 얻기 위해 AI는 프로세스와 데이터, 결과물 측면에서 안전해야 하며 적대적인 데이터와 공격에 대해 견고하게 설계되어야 함
설명가능한 AI (eXplainable AI)	• 시스템이 특정한 방식으로 행동한 이유와 특정한 해석을 내놓은 이유를 이해하는 기술로 설명 가능한 인공지능 기술 개발
시험과 검증	• 시스템의 시험과 검증은 가능한 초기부터 시행해 시스템이 수명 주기 전체에 걸쳐, 특히 배포 이후 의도한 대로 행동하도록 해야 하며 데이터, 사전 학습 모델, 환경, 전체적인 시스템의 행동을 비롯해 AI 시스템의 모든 구성요소들이 포함되어야 함

- (비기술적 방법) 규제, 행동강령, 표준화, 인증, 거버넌스, 교육, 사회적 논의, 포용적 설계팀 등 활용해 AI 윤리 요구사항 준수

비기술적 방법	내용
규제	• AI 신뢰성을 담보하는 규제 (예, 제품 안전 법률, 사용자 보호 조치 등)
행동강령	• 기업 책임 헌장, 핵심 성과 지표(KPI), 행동 강령 또는 내부 정책에 신뢰할 수 있는 AI를 위한 노력을 추가
표준화	• 설계, 제조, 사업 행위 등에 대한 표준은 구매 결정을 통해 윤리적 행동을 인식하고 독려고 품질 관리 기능을 제공 • 현재 ISO 표준 또는 IEEE P7000 표준과 같은 예가 있으며, 앞으로는 구체적인 기술적 기준에 비추어 시스템이 안전성, 기술적 견고성, 투명성을 준수한다는 것을 확인하는 “신뢰할 수 있는 AI” 표식이 적합하게 될 수도 있음
인증	• AI 시스템의 투명성과 책임성, 공정성을 대중에게 증명할 수 있는 조직을 고려 • 인증은 책임성을 대체할 수 없으므로 책임의 한계와 검토, 구제 메커니즘을 포함한 책임성의 체계를 통해 인증의 한계를 보완
거버넌스	• 내부 및 외부 거버넌스 체계를 수립해 AI 시스템의 개발과 배포, 이용과 관련 있는 의사결정의 윤리적 차원에 대한 책임성을 담보 • AI 시스템과 관련된 윤리적 사안에 대한 담당자를 지정하는 것 또는 내외부 윤리 위원회를 설치하는 것 등이 포함
교육과 인식	• 의사소통과 교육, 훈련은 AI 시스템의 잠재적 영향력에 대한 지식이 널리 퍼질 수 있게 하고 사람들이 사회적 발전을 형성하는 데에 참여할 수 있음을 인식하게 하는 데에 중요
사회적 논의	• 법률가, 기술전문가, 윤리학자, 소비자 단체, 노동자 등 다양한 이해관계자들과 함께 AI 시스템의 이용과 데이터 분석에 대해 논의
포용적 설계팀	• 설계, 개발, 시험, 유지, 배포, 조달하는 팀이 사용자와 사회 전반의 다양성을 반영할 필요

2. 신뢰할 만한 AI 윤리 평가 목록 (Self Assessment List)

□ (목적) AI-HLEG는 AI 개발 시 설계에서 출시에 이르는 전 과정에서 윤리 이슈를 자체 점검할 수 있는 평가 목록(assessment list)을 제안함

□ (내용) 기본적으로 인간의 기본권과 상충 여부 점검 그리고 신뢰할 만한 AI의 7대 윤리 요구 사항 준수와 관련된 내용들로 작성됨

□ (추진경과) 전문가 그룹이 작성한 초안('19.2)을 바탕으로 공개 파일럿 테스트를 수행하고 이해관계자들이 제안한 350건의 의견을 반영해 완성함('20.7)

○ ('19.2) 신뢰할 만한 AI 윤리 가이드라인과 함께 평가 목록 초안을 공개함

○ ('19.6~'19.12) 평가 목록 초안 완료 후 약 5개월 간('19.6.26~'19.12.1)의 파일럿 테스트[9]를 진행함

- (파일럿 테스트 목적) '평가 목록'이 업계에서 적용 가능한지 실효성을 파악하는데 주안점을 둠
- (파일럿 테스트 방법)* 온라인을 통한 설문 조사 및 우수 사례 공유, 심층 인터뷰(전화) 등 다양한 방식으로 참여자의 의견을 수렴함

* 파일럿 테스트 참여자를 대상으로 온라인 설문 조사 실시, 유럽AI위원회를 통해 모범사례 공유[10], 참여자들 중 일부를 대상으로 유선 심층인터뷰 실시 등 350여개의 이해관계자들로부터 의견 수렴

○ ('20.7) 최종 평가 목록을 온라인*으로 공개함 ('20.7.17)

* 개인 및 기관이 스스로 점검을 해 볼 수 있도록 웹 기반 툴(ALTAI)도 공개[11]

□ (항목 구성) 인간 기본권과 AI 윤리 7대 요구사항 관련 140여 가지 질문 포함

○ AI시스템의 전 개발·배포 단계에 걸쳐 적용되는 AI 윤리 요구 사항*들의 준수 여부에 대한 객관식·주관식 질문들로 구성함

9) https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/register-piloting-process-0

10) https://ec.europa.eu/futurium/en/eu-ai-alliance/BestPractices

11) https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

* (7대 요구사항) 인간행위자와 감독, 기술적 견고성과 안정성, 프라이버시·데이터 관리, 투명성, 다양성, 차별금지, 공정성, 사회·환경적 웰빙, 책임성

- 평가 목록은 AI 시스템의 상황에 적합하게 맞춤화하여 사용하길 권고함

① (기본권 보장) 인간의 기본권에 대한 영향 평가를 수행하고, 상충되는 원칙과 권리 간 절충 관계를 확인하고 문서화했는지 확인함

② (인간 행위자와 감독) 사람과의 상호작용, 통제·감독 수준이 적절한지 확인하고, 시스템 통제 메커니즘과 필요할 때 안전한 중단 수단·절차를 구비했는지 확인함

③ (기술적 견고성과 안전성) 취약성·위험을 검토하고 회복성·안전성의 보장, 위험 지표·수준의 관리와 식별·검토 방법 구비, 데이터 품질, 시스템 최신성·정확성 확보, 신뢰성·재현성 보장 여부를 확인함

④ (프라이버시와 데이터 거버넌스) 민감한 데이터 최소화, 개인정보보호 강화 조치 여부를 확인하고, 데이터 품질·무결성 감독 절차를 구비하고 표준을 준수하는지 검토함

⑤ (투명성) 개발 적용 방식과 결과 활용에 대한 의사결정사항이 추적 가능한지, 자동화된 의사 결정이 설명가능한지, 시스템의 목적·용례·원리·한계 등을 이해관계자에게 고지했는지 확인함

⑥ (다양성, 차별 금지, 공정성) 공정성을 적절히 정의하고 편향 방지 절차를 구비·고지했는지 확인하고, 보편적 설계와 접근을 보장하는지, 설계·개발 과정에 이해관계자가 참여하는지 확인함

⑦ (사회적, 환경적 웰빙) 지속 가능성과 친환경성, 일자리 및 직무에 미치는 영향 등 사회적 영향력을 평가했는지 확인함

⑧ (책임성) 문서화 등을 통한 추적가능성을 보장, 감사 가능성 확인, 책임성 관련 교육·훈련 여부와 AI 윤리 검토 위원회 설치, 위험 보고 절차 수립 등 위험관리 절차 및 지침이 마련되어 있는지 검토함

[표3-2] EU 신뢰할만한 AI 윤리 가이드라인 자가 평가 목록

카테고리(문항 수)		평가 내용
기본권 보장 (12)		인간의 기본권에 대한 영향 평가를 수행하고, 상충되는 원칙과 권리 간 절충 관계를 확인하고 문서화함
인간행위자와 감독	인간행위자와 자율성 (14)	의사결정의 자동화 수준을 고려할 때 사람과의 상호작용과 통제·감독 수준이 적절한지 확인
	인간의 감독 (8)	HITL(Human-in-the-loop), HOTL(Human-on-the-loop), HIC (human-in-command)와 같은 시스템 통제 메커니즘의 확보 여부, 필요할 때 안전한 중단 수단·절차를 구비했는지 확인
기술적 견고성과 안정성	공격회복성과 보안(9)	시스템의 취약성과 잠재적 위험을 검토하고 대비하여 회복성과 안전성을 보장하기 위한 대비 계획·절차·보험·거버넌스 등 구비 여부 확인
	일반적 안전성(10)	위험 정의, 위험 지표 및 수준 관리, 잠재적 위협 식별 및 예상 결과 검토 여부, 핵심 시스템의 의존성 등 평가 방법 확보 여부
	정확성(5)	데이터의 품질, 시스템의 최신성, 정확성 확보 노력 등 평가
	신뢰성, 대체계획, 재현성(9)	신뢰성과 재현성의 중요성, 신뢰성 및 재현성의 보장 방법, 시스템 에러에 대비한 대체 계획의 여부
프라이버시·데이터 관리	프라이버시(2)	민감한 데이터를 최소화하고 적절한 개인정보보호 강화조치 여부 확인
	데이터 거버넌스(11)	데이터의 품질 및 무결성 감독 절차를 수립 여부, 국제 관리 표준을 준수 여부 확인
투명성	추적가능성(6)	시스템을 설계·개발·테스트·검증하는데 사용되는 방법과 결과 활용에 대한 의사 결정 사항들을 문서화하여 추적이 가능한지 확인함
	설명가능성(2)	시스템에 의한 선택이나 자동화된 의사 결정이 설명이 가능한지 확인함
	고지의 의무(5)	시스템의 목적·용례·특성·한계를 이해관계자에게 충분히 고지했는지 확인함
다양성, 차별금지, 공정성	편향 회피(15)	시스템 설계 시 사용한 공정성의 정의가 목적, 사용자 등에 적절한 수준인지, 불공정한 편향을 지양하기 위한 절차와 메커니즘을 수립하고 투명하게 고지했는지 확인함
	접근가능성·보편적 설계(9)	접근성·이해관계자의 다양성을 보장하고 보편적인 디자인을 채택했는지 확인함
	이해관계자 참여(1)	시스템의 설계 및 개발 과정에서 다양한 이해관계자의 참여 기회 여부
사회·환경적 복지	환경적 영향(4)	지속 가능하고 환경 친화적인 방식의 시스템 구현 및 운영 여부
	일자리 및 직무영향(8)	시스템이 근로 환경, 고용 관계, 직무 등에 미치는 영향을 평가
	거시적 사회 영향 및 민주주의(4)	제도, 민주주의, 거시적 차원에서의 시스템의 사회적 영향력 평가
책임성	감사가능성(2)	시스템의 절차 및 결과물에 대한 감사 기능을 적절하게 구현하여 추적성과 감사가능성을 보장하고 있는지, 상충되는 원칙 간 절충에 대한 의사결정이 문서화 되었는지 확인함
	위험관리(10)	시스템의 영향 평가와 책임성 강화 교육·훈련을 하고 있는지, AI 윤리 검토 위원회를 설치하고 내부 지침을 마련했는지, 제 3자나 근로자에게 위험 보고 프로세스를 수립했는지 확인함

* 출처 : Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment ('20.7)

Ⅳ. 시사점

□ (체계적 정책 개발) EU의 AI 윤리 가이드라인 수립은 EU차원의 AI 전략 발표 이후 2년의 시간을 통해 투명하고 체계적으로 진행되어 왔음

○ (장기적 추진) AI 기술의 사회적 통합은 장기적 비전하에 추진하는 것이 바람직

- AI의 사회적 영향을 다각도에서 충분히 검토하고 이에 대응하는 법·제도 정비*, 대응 기술 개발은 장기적 추진**이 권장됨

* 영국 정부가 최근 발간한 『AI로드맵』('21.1)은 AI기술, 윤리 및 보안, 사회적 영향 등을 고려해 향후 10-50년의 장기 비전을 가지고 정책을 추진할 것을 권고함[12)]

** 스탠포드大는 향후 100년 간의 AI 연구를 정기적으로 점검하겠다는 AI100 추진('16.9)

○ (투명성 확보) AI 윤리 정책의 내용과 보고서 작성 과정에서 투명성을 확보 함

- 주요 보고서 발간 시 다양한 이해관계자로 구성된 협의체를 통해 공개적으로 의견을 수렴해 수정본에 반영했으며 수집된 의견도 온라인으로 공개

- 우리나라도 윤리 인식 확산과 투명성 확보 측면에서 향후 구체적인 가이드라인 또는 체크리스트의 개발 시 의견 수렴 과정과 내용의 공개가 바람직*

* 우리나라는 지난 '19년 AI 국가전략 발표 후 '20년 12월 국가 AI윤리 기준을 수립하고 관련 법 정비와 AI윤리 체크리스트 개발을 진행하는 단계임

○ (실효성 제고) 가이드라인 및 체크리스트는 실증 사업을 통해 실무 적용 과정에서의 애로사항을 반영해 실효성을 제고해야 함

- EU의 가이드라인은 다양한 이해관계자의 의견을 바탕으로 작성되었으나 실무현장에서의 실증은 되지 않음

- 향후 국내 AI 윤리 가이드라인 또는 체크리스트 개발 시 선행 사례*를 종합적으로 고려하고 기업 현장에서의 실증 조사를 통해 실효성을 검증해야 함

* 영국 정보위원회 AI감리 프레임워크('20.2), 마이크로소프트 AI 공정성 체크리스트('20.4), 카네기멜론大 윤리적 AI경험 설계 체크리스트('19.12) 등

12) https://www.gov.uk/government/publications/ai-roadmap/executive-summary

□ (자율적 실천) EU의 가이드라인은 AI 시스템 개발 주체의 자율적 실천을 바탕으로 하며 궁극적 목적은 AI 개발 주체의 '책임 있는 경쟁력' 강화에 있음

○ (법과 윤리의 구분) EU 가이드라인 초안의 공개 의견 수렴 과정에서 법과 윤리의 명확한 개념적 구분과 사업자들이 책임을 져야하는 구체적 법률 정보에 대한 정보 요청이 다수 있었음

- EU 가이드라인은 개발자들이 어떻게 하면 AI기술을 윤리적*으로 사용할 수 있을지 구체적 지침을 줄 목적으로 작성되어 구속력은 없음

  * AI로 '무엇을 할 수 있는(can do)'보다 '무엇을 해야 하는지(should do)'에 대한 지침

○ (모범 사례 확산) 대기업 및 선도 기업들이 자율적으로 실천하는 모범 사례*를 발굴하여 중소기업, 스타트업에 보급·확산하는 정책이 필요함

  * 카카오 AI 윤리 헌장('18.1) 및 전직원 대상 AI 윤리 교육('20.2), 네이버-서울대 AI 윤리 준칙('21.2) 등의 현장 실천 사례 참고

○ (유연한 맞춤화) 자율 점검을 위한 AI 윤리 평가 목록은 AI 시스템의 목적과 상황에 맞춰 자유롭게 수정 보완해 적용할 것을 권고하고 있음

- 기계학습 결과의 편향성 및 학습 오류의 완벽한 사전적 예방은 기술적으로 매우 어려움

  * 윤리적 AI 시스템 개발에 △ 윤리 항목 간의 충돌, △ 데이터와 AI 모델 간극의 모호성, △ 공정성과 성능의 반비례 관계 등 기술적 난제가 따름 (Nature, '20)[13]

- 기업은 윤리적으로 불완전한 AI 제품 및 서비스의 출시로 인한 사회적 비판과 매출 불이익 등의 피해를 최소화할 유인이 있음

○ (신중한 입법) 기본권 침해, 보안, 프라이버시 보호 등 기존 법규제의 공백을 분석하되 산업 혁신을 고려해 규제를 위한 입법은 신중한 접근이 필요함

- 가이드라인은 윤리적 AI 시스템의 개발을 지원하기 위한 것으로 기업의 사회 책임을 실현하는 도구로서 역할을 할 뿐 구속력을 갖지 않음

13) https://www.nature.com/articles/s41599-020-0501-9

- EU는 AI에 의한 기본권 침해와 보안 이슈 관련 새로운 규제 입법이 필요* 하나 산업계의 혁신 친화적 정책에 대한 요청과 균형점을 모색 중[14)]

  * EU는 AI 입법 영향 평가 마쳤고 '21년 1분기 내에 관련 입법 제안서 제출 예정[15)]

- 국내의 개인정보 관련법*, 데이터 3법 등 관련법의 이해가 부족한 상황에서 AI윤리 가이드라인이 규제로 작용할 경우 AI사업의 법적 리스크 상승 우려

  * (예) 개인 의료정보의 경우 개인정보보호법에서는 '민감정보', 보건의료기본법에서는 '보건의료정보', 의료법은 '의무기록', '진료정보' 등 용어나 개념이 혼재[16)]

□ (지속적 국제 동향 파악) 향후 EU의 AI 입법 동향을 비롯해 AI윤리 관련 주요 기업 동향, 기술 표준화 등 국제 사회 논의를 지속적으로 파악할 필요가 있음

○ (국제 협의체) OECD, GPAI(Global Partnership on Artificial Intelligence) 등 국제 사회의 논의에 참여를 강화하여 국내 AI 윤리 규범을 지속적인 보완이 필요함

- OECD의 AI원칙('19.5)은 36개 회원국과 6개 파트너국이 채택한 최초의 정부간 표준 협약이며 G20 AI 원칙으로도 채택 됨

- GPAI는 2020년 6월 G7을 비롯해 우리나라를 포함 15개 국이 결성한 AI 협의체로 AI R&D 지원, 미래 일자리, 책임성 있는 AI 등 이슈 논의 중

○ (국제 표준 단체) 국제 표준 단체(ISO, IEEE 등)에서의 AI 윤리 표준 활동도 지속적 지원이 요구됨

- AI 국제 표준은 ISO, IEC, IEEE 등에서 관련 분과 및 실무 그룹을 신설해 2017년부터 AI윤리 관련 표준 논의를 해오고 있으며 최근 본격화

14) Lucilla Sioli, DIRECTOR OF ARTIFICIAL INTELLIGENCE AND DIGITAL INDUSTRY, DG CONNECT, (2020.12.23.), Two years of policy reflection on AI: our way forward, https://ec.europa.eu/digital-single-market/en/blogposts/two-years-policy-reflection-ai-our-way-forward

15) EC(2020.7.23.), Proposal for a legal act of the European Parliament and the Council laying down (Ref. Ares(2020)3896535 - 23/07/2020) requirements for Artificial Intelligence, https://ec.europa.eu/info/law/better-regulation/

16) 4차산업혁명위원회, 국가 데이터 정책 추진 방향, 2021.2.17

* (관련단체) △(ISO/IEC JTC1) 분과위원회 42(SC42) △(IEC SEG10) 자율 및 AI 애플리케이션 윤리 분과 △(IEEE SA[17]) 전기전자 관련 13개 실무 그룹 운영 중

- 우리나라는 국립전파연구원(과기정통부)과 국가기술표준원(산업통상자원부)에서 인공지능 국제 표준화 공동 대응 계획을 발표('20.11)

* (국립전파연구원) TTA-인공지능기반기술(PG1005) 프로젝트그룹에서 AI표준 추진 (국가기술표준원) SC42 국제표준 대응 및 AI윤리 표준화 포럼 논의 중

○ (비영리 단체) OpenAI, Partnership on AI, Future of Life Institute (FLI) 등 AI 관련 주요 비영리 기관의 활동 동향 파악

* (OpenAI) GPT-3 개발 시 데이터 및 알고리즘의 편향 제거 노력 및 오남용 금지의 선언적 명시, (Partnership on AI) 美 글로벌 IT 기업 중심의 비영리단체 (FLI) AI 개발 원칙(아실로마 AI 원칙) 제정

17) https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

[별첨] AI윤리 가이드라인 초안('18.12.18) 의견 수렴 결과

문서	의견(contribution)	반영사항
서론	• AI에 대한 지나친 비관론적 접근 지적	• 균형 잡힌 어조, 구어체 등의 생략
	• 문서 간략화 필요, 챕터 통합 제안	• 챕터는 유지했으나 챕터간의 유기적 연결을 위한 설명 강화
	• 가이드라인의 승인 메커니즘과 가이드라인을 준수하지 않을 경우 잠재적 결과에 대한 설명 요구	• 가이드라인의 승인 절차는 현재로서는 없음 • 가이드라인은 자발적 사용 기반임을 명시
	• 모든 AI 애플리케이션에 적용되는 가이드라인보다 섹터별 맞춤화 방안에 대한 접근 필요	• 파일럿 테스팅을 통해 섹터별 특성을 파악하여 가이드라인 및 평가리스트 수정 계획
	• AI실무자들과 이해관계자들에 대한 다양한 설명 필요	• Ch2의 신뢰할만한 AI 요구사항에 AI 시스템 관련 다양한 이해관계자의 역할 설명
	• 가이드라인의 '지리적 범위' 확장 필요, 유럽에서 만든 AI 시스템에만 적용될 것이 아니라 타국에서 만든 것도 유럽에서 배포, 사용될 경우에도 적용되어야 함	• 유럽 외 국가에서 개발된 AI 시스템도 가이드라인의 권고를 따를 것을 명시하는 '유럽을 위한 신뢰할 만한 AI'라는 방향으로 수정
	• 용어 설명 추가, 가이드라인에 해당자 명시	• 추가적인 용어 설명, 부제목에 가이드라인의 해당자와 범위 기재, 다른 문서에서 변경된 AI의 정의를 병행적으로 수정 반영
	• AI가 적용받는 법률 목록 요청	• 법률 목록이 방대하고 불완전할 가능성이 있어 이를 전부 제공하지는 않으나 소론에 '적법한 AI'관련 내용에 몇 가지 관련 법률 예시를 제공
	• AI에 대한 설명과 일반적인 컴퓨팅에 대한 설명을 명확히 구분할 것	• 수정본에도 AI에만 관련된 설명과 일반적인 컴퓨팅에 대한 설명을 구체적으로 구분하지는 않고 있으며 이러한 구분이 쉬운 것도 아니고 특정 측면을 간과할 위험도 있음 • AI HLEG는 AI와 관련이 있다고 생각하는 문제 해결에 초점을 맞춤
Ch1 신뢰할 만한 AI의 기초	• 윤리와 법의 구분이 모호, 둘 사이의 관계를 명확히 할 필요	• '신뢰할 만한 AI'를 정의(합법성, 윤리성, 견고성)하는 방식으로 법과 윤리를 보다 명확히 구분 • 가이드라인을 법적 의무를 준수하는 방법에 대해 조언할 의도가 없음을 명시
	• 사전 동의, 옵트 아웃 권리와 같은 GDPR 관련 용어와의 혼동을 피해야 함	• GDPR과의 용어 정합성 검토 후 수정
	• 두 가지 이상의 원칙이 상충되는 상황을 해결하는 방법에 대한 명확성을 요청	• 수정본의 2.3항에 이러한 윤리 간 상충 상황에 대한 대응 방법을 설명 • 절충 사항에 대한 결정의 문서화와 책임 소재의 확인 필요성 강조
	• 선행 원칙 (do good)과 같은 도덕적 의무는 원칙이 아니며 문맥에 적합지 않음	• '선행 원칙'에 대한 언급 삭제
	• AI 근로자 또는 노동에 미치는 영향에 대한	• 초안에 이미 근로자 간의 힘이나 정보의

	언급이 누락되었다고 지적	비대칭성을 특징으로 하는 상황에 주의를 기우릴 것을 지적하고, 작업 환경에 대해서는 인간 자율성 존중 원칙을 언급
CH2 신뢰할 만한 AI의 실현	• ch1의 윤리원칙과 ch2의 명시적 요구 사항 사이의 연관성을 명확히 설명할 것을 요청	• 기본권에 기초한 윤리 원칙들로부터 7가지 요구사항이 도출된 것으로 설명 • 각 요구사항이 어떤 윤리원칙들과 연결되어 있는지 설명
	• 특정 요구사항을 다른 것과 합치고 또 좀 더 자세히 설명해 줄 것으로 요청	• 이전 10개의 요구사항 대신 7개의 요구사항으로 병합했으며 각 요구 사항은 세부적인 제목으로 여러 측면으로 설명[18]
	• 환경과 다른 생명체에 대한 관심 부족 지적	• Societal and Environmental Wellbeing을 요구사항의 하나로 반영
CH3 신뢰할 만한 AI의 평가	• 머리말, 질문 등이 중복 지적	• 평가 목록을 더 명확히 하고 반복적인 표현을 삭제
	• 매우 다양한 환경에서 모든 AI 애플리케이션을 포괄할 수 있는 평가 목록의 가능성에 대한 언급	• 수정 지침에서는 평가 목록이 시스템이 작동하는 특정 사용 사례와 상황에 맞춰 조정되어야 한다고 명시함 • 이는 AI개발자, 배포자, 사용자가 직접 수행해야 함 • 파일럿 테스트를 통해 횡적으로 적용될 수 있는 부분과 잠재적으로 보완될 필요가 있는 부분을 식별
	• 평가 목록에 대한 준수 여부를 평가하기 위한 지표 및 KPI에 대한 지침 제공 또는 인증 라벨 부여 등에 대한 고려 필요	• 이 가이드라인은 구속력 없는 지침임 • 평가 목록을 준수하더라도 '신뢰할 수 있는 AI'라는 인증 라벨을 부여할 수 없음 • 지침의 준수 평가를 위한 측정 항목이나 KPI는 자유롭게 개발 가능
	• 평가 목록에 있는 질문에 대한 답안 요청	• 지침의 평가 목록에 있는 질문에 대한 별도의 답을 제공하지 않음 • 지침의 목적은 신뢰할 수 있는 AI가 어떻게 운영될 수 있는지에 대한 성찰을 장려하는 것을 목표로 작성됨 • 파일럿 테스트 단계에서 이해관계자간의 모범 사례를 교환하도록 장려할 예정

18) Accountability, Data Governance, Design for All, Governance of AI Autonomy (Human Oversight), Non-discrimination, Respect for (& enhancement) Human Autonomy, Respect for Privacy, Robustness, Safety, Transparency 이상 10가지 → Human Agency and Oversight, Technical Robustness and Safety, Privacy and data governance, Transparency, Diversity non-discrimination and fairness, societal and environmental wellbeing, accountability 7가지로 수정

참고문헌

1. 참고문헌

- 관계부처합동(2019.12), 인공지능 국가 전략
- 산업통상자원부(2020.11), 산업부·과기정통부 힘을 모아 인공지능 국제표준화 공동대응 전략 모색
- 4차산업혁명위원회, 국가 데이터 정책 추진 방향, 2021.2.17
- KPC4IR(2019.12), 인공지능의 윤리/정책/사회 이슈, ISSUE PAPER No. 08

2. 국외문헌

- AI-HLEG(2019.4.8.), A definition of Artificial Intelligence: main capabilities and scientific disciplines
- European Commission (2018.4.25), Communication Artificial Intelligence for Europe
- European Commission (2018.12.7), Coordinated Plan on Artificial Intelligence
- European Commission (2020.7.17), Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
- European Commission (2020.7.23.), Proposal for a legal act of the European Parliament and the Council laying down (Ref. Ares(2020)3896535 - 23/07/2020) requirements for Artificial Intelligence
- Gov.uk (2021.1.6), AI Roadmap
- Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety. The Alan Turing Institute
- Samuele Lo Piano (2020.6.17.), Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward, Nature : Humanities and Social Sciences Communications, Vol 7, Article number: 9.

3. 기 타

- High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence, 최종접속 2020.2.24
- European Commission (2019.6.26), The first European AI Alliance Assembly, https://ec.europa.eu/digital-single-market/en/news/first-european-ai-alliance-a

ssembly, 최종접속 2021.2.24
- https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines, 최종접속 2021.2.24
- European Commission, Pilot the Assessment List of the Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/register-piloting-process-0, 최종접속 2021.2.24
- European Commission, Piloting the Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/futurium/en/eu-ai-alliance/BestPractices, 최종접속 2021.2.24
- European AI Alliance, ALTAI - The Assessment List on Trustworthy Artificial Intelligence, https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence, 최종접속 2021.2.24.
- European Commission, ETHICS GUIDELINES FOR TRUSTWORTHY AI - OVERVIEW OF THE MAIN COMMENTS RECEIVED THROUGH THE OPEN CONSULTATION -, https://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/stakeholder-consultation-guidelines-first-draft, 최종접속 2021.2.24.
- IEEE SA, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html, 최종접속 2021.2.24.
- Lucilla Sioli, DIRECTOR OF ARTIFICIAL INTELLIGENCE AND DIGITAL INDUSTRY, DG CONNECT, (2020.12.23.), Two years of policy reflection on AI: our way forward, https://ec.europa.eu/digital-single-market/en/blogposts/two-years-policy-reflection-ai-our-way-forward, 최종접속 2021.2.24.
- Stanford University, One Hundred Year Study on Artificial Intelligence (AI100), https://ai100.stanford.edu/, 최종접속 2021.2.24

주 의

이 보고서는 소프트웨어정책연구소에서 수행한 연구보고서입니다.
이 보고서의 내용을 발표할 때에는 반드시
소프트웨어정책연구소에서 수행한 연구결과임을 밝혀야 합니다.

유럽(EU)의 인공지능 윤리 정책 현황과 시사점 : 원칙에서 실천으로

An European approach to ethics on artificial intelligence: from principles to practices

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D 연구동(A)
Global R&D Center 4F 22 Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do

www.spri.kr

ISSN 2733-6336