TTA Technical Report

기술보고서

TTAR-xx.xxxx

제정일: 2019년 09월 09일

인공지능 윤리 가이드라인 제정 동향(기술보고서)

Enactment Trends on Artificial Intelligence Ethics Guidelines(Technical Report)

TTA 한국정보통신기술협회
Telecommunications Technology Association

기술보고서 초안 검토 위원회 인공지능기반기술 프로젝트그룹(PG1005)

기술보고서안 심의 위원회 지능정보기반 기술위원회(TC10)

	성명	소 속	직위	위원회 및 직위	기술보고서번호
기술보고서(과제) 제안	오재섭	숙명여대경영전문대학원	대우교수	-	
기술보고서 초안 작성자	오재섭	숙명여대경영전문대학원	대우교수	-	
사무국 담당	박예슬	TTA	선임	-	

본 기술보고서 발간 이전에 접수된 지식재산권 확약서 정보는 본 기술보고서의 '부록(지식재산권 확약서 정보)'에 명시하고 있으며, 이후 접수된 지식재산권 확약서는 TTA 웹사이트에서 확인할 수 있습니다.
본 기술보고서와 관련하여 접수된 확약서 외의 지식재산권이 존재할 수 있습니다.

발행인 : 한국정보통신기술협회 회장
발행처 : 한국정보통신기술협회
13591, 경기도 성남시 분당구 분당로 47
Tel : 031-724-0114, Fax : 031-724-0109
발행일 : 20xx.xx

서 문

1 기술보고서의 목적

이 기술보고서의 목적은 EU, IEEE, ACM, Internet Society, Google 등이 제정한 인공지능 윤리 가이드라인을 인공지능 시스템 디자인-개발-서비스 제공 관점으로 정리하여 인공지능 분야의 서비스 제공자에게 인공지능 관련 윤리에 대한 전반적인 가이드라인 참고 자료를 제시하여, 인공지능 시스템 디자인-개발-서비스 제공에 윤리적 고려사항을 우선시할 수 있도록 하는 정보를 제공하는 것에 있다.
인공지능 기술이 확산될수록 인공지능 기술의 사회적 수용을 위태롭게 하고, 인공지능 기술의 사회적 수용을 방해하는 윤리적이고 도덕적인 문제에 직면하고 있다. 인공지능 시스템의 자율성이 높아지고 물리적 또는 디지털 환경에 대한 통제력이 높아짐에 따라 의도하지 않은 결과가 발생하면서 오류가 발생할 가능성이 높아지고 있다. 이것은 의도하지 않게 위해를 가할 수 있는 인공지능 기술의 사고를 의미하며, 일반적으로 기계에 적절한 객관적인 기능과 훈련 데이터가 제공되지 않을 때 발생한다.

과학자들은 인공지능 학습 프로세스의 이러한 측면을 완전히 제어하고 인공지능 학습 시스템의 모든 단일 행동을 정확하게 예측할 수 있는 위치에 있지 않다. 그러나 현재 인공지능을 보다 안전하고 안정적으로 사용하기 위해 이러한 사고를 최소화하고 견고성, 위험 감도 및 안전한 탐색을 개선하기 위해 노력하고 있다.

그래서 인공지능 기술과 시스템에 의해 제기된 윤리 문제는 보다 광범위하게 분석되어야 하며 이에 따라 적절한 윤리 정책을 개발해야 한다. 이는 일관되고 신뢰할 수 있는 규제 프레임워크를 만들어야 한다는 것을 의미한다. 이는 보다 효과적이고 비용 효율적인 것으로 간주되는 특정 알고리즘이 아닌 광범위한 인공지능 원칙을 규제함으로써 달성될 수 있을 것으로 판단된다. 이 보편적인 윤리적 프레임워크는 인간 중심의 인공지능 원칙(즉, 인간의 감독과 통제 하에 긴밀히 개발된 인공지능)에서 영감을 얻어 인공지능 시스템의 설계, 구현 및 테스트의 모든 단계에 적용되는 알고리즘에 공정성과 정의의 원칙을 내장시킬 수 있어야 할 것이다.

2 주요 내용 요약

이 기술보고서는 인공지능 서비스 제공자에게 인공지능 관련 윤리에 대한 전반적인 가이드라인 참고 자료를 제시하여, 인공지능 서비스 제공자가 인공지능 시스템을 디자인하고 개발하고 서비스하는 경우에 이 기술보고서를 참조하여 윤리적 고려사항을 우선시할 수 있도록 참고 자료를 제공하는 것에 있다.
'EU의 믿을만한 인공지능을 위한 윤리 가이드라인'은 인공지능 시스템의 신뢰를 구축하는 방식으로 개발, 배포 및 이용되는 것을 보장하기 위해 기본권에 대한 4 가지 윤리 원

칙(인간자율성 존중, 피해 방지, 공정성, 설명가능성)을 제시하였다. 또한 신뢰할 수 있는 인공지능 구현을 위한 요건을 제시하였다. 자율성과 감독, 기술적 견고함과 안전성, 프라이버시와 데이터 거버넌스, 투명성, 다양성/비차별성/공평성, 사회적 환경적 복지, 책임성을 요건으로 제시하였다. 4가지 원칙은 기존 법적 요구 사항에 이미 상당 부분 반영되어 있으므로 윤리 원칙을 준수하는 것은 기존 법을 공식적으로 준수하는 것 이상의 의미를 부여하는 인공지능 윤리 가이드라인이다.

'IEEE의 윤리적으로 조율된 설계(EAD)'는 지능형 제품, 서비스 개발 관련 윤리적 설계 가이드라인을 제시하기 위해 일반 원칙 5가지(인권, 복리, 책무성, 투명성, 오남용에 대한 인식)를 정하고. 자동화된 시스템에 응용된 기술들에 윤리적 고려사항을 반영하고 이를 측정할 표준을 제정하는데 초점을 두고 있다. 특히, 윤리적 가치를 자율지능 시스템에 내장시키기 부분은 인공지능 시스템 개발과 연관성이 매우 높다.

'ACM의 윤리강령과 전문가 행동 강령'은 강화된 윤리 강령 및 전문가 행동 강령에서 인공지능 윤리와 연관있는 알고리즘 투명성과 책무성 관련 가이드라인을 제시하였다. 알고리즘 투명성과 책무성에 대한 성명에서 7가지를 제시하였다. 7가지는 인식, 접근과 수정, 책무성, 설명가능성, 데이터 출처 파악, 감가가능성, 유효성 검사와 테스트이다. 2018년 개정 윤리 강령에서 인공지능 윤리 제시. 인공지능 윤리 항목으로 피해의 회피, 투명성과 공정성, 차별에 대한 주의, 개인정보의 존중, 대중인식, 안전을 제시하였고 항목별로 행동 강령을 제시하였다.

'Internet Society의 인공지능과 기계학습'은 인공지능 시스템의 설계와 배포에서 원칙, 윤리적 고려사항과 권고 사항으로 구성되어 있다. 권고사항은 인간의 통제, 안전 최우선, 개인 정보 보호. 데이터 공급의 신중성, 인공지능 시스템의 안전성, 책임 배포이다.

'아실로마 인공지능 23원칙'은 아실로마 인공지능 23가지 원칙은 연구 이슈, 윤리 및 가치, 장기 이슈 3 부분으로 구성되어 있다. 3부분 가운데 윤리와 가치(Ethics and Values)와 직결된 항목들은 13가지이다. 16가지 항목은 안전, 장애 투명성, 사법적 투명성, 책임, 가치 정렬, 개인정보 보호, 자유와 개인 정보, 공동 이익, 공동 번영, 인간의 통제력, 비파괴, 인공지능 무기 경쟁이다.

'구글 인공지능 원칙'은 인공지능 관련 비즈니스 의사결정에 영향을 주는 구체적인 기준 7가지를 제시하였다. 7가지 기준은 개인정보보호, 책임, 안전, 편견 조장 금지, 인간의 지시와 통제, 높은 수준의 과학적 수준 유지이다.

본 보고서에서 다루는 내용을 요약하면 다음과 같다.

- 모든 유형의 인공지능에 적용할 수 있는 최고수준의 윤리 원칙
- 인공지능 윤리 가이드라인과 법 제도와의 관계
- 지능형 제품, 서비스 개발 관련 윤리적 설계 가이드라인을 제시하기 위해 일반 원칙 5가지(인권, 복리, 책무성, 투명성, 오남용에 대한 인식)
- 강화된 윤리 강령 및 전문가 행동 강령에서 인공지능 윤리와 연관있는 알고리즘 투명성과 책무성 관련 가이드라인
- 인공지능 시스템의 설계와 배포에서 원칙, 윤리적 고려사항과 권고 사항

3 인용 기술보고서와의 비교

3.1 인용 기술보고서와의 관련성

해당사항 없음

3.2 인용 표준과 본 기술보고서의 비교표

해당사항 없음

Preface

1 Purpose

The purpose of this technical report is to summarize AI ethics guidelines established by EU, IEEE, ACM, Internet Society, Google, etc. from the perspective of AI system design, development, and service provision. The overall guideline for this document is to provide information to prioritize ethical considerations in AI system design-development-service delivery.

2 Summary

This technical report is intended to provide AI service providers with a comprehensive guideline on AI-related ethics. In particular, in the case of AI service providers designing, developing and servicing AI systems, this Technical Report provides a reference for prioritizing ethical considerations.

The summary of this report is as follows.

- The highest ethical principles applicable to all types of AI Relationship between AI Ethics Guidelines and Legal System
- Five general principles (recognition of human rights, benefits, accountability, transparency, and abuse) to present ethical design guidelines for developing intelligent products and services.
- Guidelines on algorithm transparency and accountability in relation to AI ethics in the enhanced Code of Ethics and Expert Conduct
- Principles, ethical considerations and recommendations in the design and deployment of AI systems

3 Relationship to Reference Standards

None

목 차

1. 적용 범위 ····· x

2. 인용 표준 ····· x

3. 용어 정의 ····· x

4. 약어 ····· x

5. 인공지능 윤리 가이드라인 주요 동향 ····· x
   5.1 EU의 믿을만한 인공지능을 위한 윤리 가이드라인 ····· x
   5.2 IEEE의 윤리적으로 조율된 설계(EAD) ····· x
   5.3 ACM의 윤리강령과 전문가 행동 강령 ····· x
   5.4 Internet Society의 인공지능과 기계학습 ····· x
   5.5 아실로마 인공지능 23원칙 ····· x
   5.6 구글의 인공지능 원칙 ····· x
6. 인공지능 윤리 가이드라인 주요 동향 ····· x

부속서 A (자유 작성 부속서) 제목 ····· x
부록 I (자유 작성 부록) 제목 ····· x
부록 II-1 지식재산권 확약서 정보 ····· x
II-2 시험인증 관련 사항 ····· x
II-3 본 기술보고서의 연계(family) 표준 ····· x
II-4 참고 문헌 ····· x
II-5 영문기술보고서 해설서 ····· x
II-6 기술보고서의 이력 ····· x

인공지능 윤리 가이드라인 제정 동향 (Enactment Trends on Artificial Intelligence Ethics Guidelines)

1 적용 범위

이 기술 보고서는 인공지능 서비스 제공자에게 인공지능 관련 윤리에 대한 전반적인 가이드라인동향을 제시하고, 인공지능 서비스 제공자가 인공지능 시스템을 디자인하고 개발하는 경우에 이 보고서를 참조하여 윤리적 고려사항을 우선시할 수 있도록 하는 데 있다.

2 인용 표준

'해당 사항 없음'

3 용어 정의

3.1 인공지능(AI, Artificial Intelligence)

인공지능은 인간의 두뇌와 같이 컴퓨터 스스로 추론, 학습, 판단하면서 전문적인 작업을 하거나 인간 고유의 지식 활동을 하는 시스템을 의미한다.

[출처] TTA 용어사전

3.2 자율시스템(AS, Autonomous System)

정보처리 능력을 보유한 인공물이 인공물 자신의 경험에 의해 행동의 방향을 결정할 수 있게 하는 입출력 및 처리 능력을 가진 시스템을 의미한다.

3.2 자율지능 시스템(AIS, Autonomous Intelligence System)

자율지능 시스템은 인간이 자신의 경험에 의해 행동의 방향을 결정할 수 있는 로봇과 상호작용에서 로봇을 지각하고, 로봇과의 상호작용을 관리할 수 있게 하는 시스템을 의미한다.

3.3 개인 식별 정보(PII, Personally Identifiable Information)

개인 식별 정보는 개인의 유일한 물리적, 디지털, 혹은 가상적 자기동일성에 기반하여 개인에게 합리적으로 연결된 모든 데이터를 의미한다.

[출처] IEEE, Ethically Aligned Design, 2019.

3.4 인공지능 윤리

인공지능 윤리는 인공지능과 인공지능 시스템이라는 인공물이 인간의 생활에서 있어 바람직한 상태, 인간과 인공지능 시스템과의 좋은 관계 설정, 인공지능 시스템 설계와 배포에서 노력할 만한 것은 무엇인지 등을 밝히는 것을 의미한다.

4 약어

AI Artificial Intelligence
AS Autonomous System
AIS Autonomous Intelligent System
PII Personally Identifiable Information

5 인공지능 윤리 가이드라인 제정 동향

5.1 EU의 믿을만한 인공지능을 위한 윤리 가이드라인

5.1.1 믿을만한 인공지능의 기초가 되는 기본권

국제인권법, EU 조약 및 EU 헌장에 명시된 포괄적 개인 권리 집합 중에서, 아래의 기본권 조항들은 인공지능 시스템에 대한 시사점을 제공한다. 특정 환경에서 다수의 권리는 EU에서 법적으로 강제되어 규정 준수가 법적으로 의무 사항이다.

그러나 법적으로 강제되는 기본권을 준수한 이 후에도, 윤리적 반영은 인공지능 시스템의 개발, 배포 및 사용이 기본권과 기본 가치에 영향을 미치는 정도를 이해하는 데 도움을 주며, (현재) 기술로 할 수 있는 것보다 우리가 해야 할 일을 확인할 때 명확한 가이드라인을 제공한다.

5.1.1.1 인간 존엄성의 존중

인간 존엄성은 모든 인간이 "본질 가치"를 지니고 있다는 아이디어에서 출발한다. 인간 존엄성은 인공지능 시스템 같은 새로운 기술에 의해 결코 약화되거나 타협되거나 억눌려서는 안 된다. 이런 맥락에서 인간 존엄성의 존중은 모든 사람이 단순히 선별되고, 분류되고, 채점되고, 갇히고, 조절되거나 조작될 수 있는 대상이 아닌, 도덕적 주체로서 모든

사람을 존경심으로 대우해야 하는 것을 의미한다. 인공지능 시스템은 인간의 육체적, 정신적 청렴성, 개인 및 문화적 정체성, 그리고 필수 요구 사항을 존중하고, 섬기고, 보호하는 방식으로 개발돼야 한다.

5.1.1.2 개인의 자유

인간은 스스로 자유롭게 삶의 의사결정을 내릴 수 있어야 한다. 이는 주권 침입으로부터 자유를 수반하지만, 배제의 위험에 처해있는 개인이나 사람들이 인공지능의 이익과 기회에 동등하게 접근할 수 있도록 정부 및 비정부 기구의 개입도 필요하다. 인공 지능의 맥락에서 개인의 자유는 (직)간접 불법적 강압, 정신 자율성과 정신 건강에 대한 위협, 정당하지 않은 감시, 속임수 및 불공정 조작에 대한 대비를 필요로 한다. 사실, 개인의 자유란 (다른 권리들 중에서도) 사업 수행의 자유 보호, 예술과 과학의 자유, 표현의 자유, 사적 삶과 사생활 보호, 그리고 집회와 결사의 자유에 대한 권리를 포함하여 개인 삶에 대한 더 높은 통제권을 행사할 수 있도록 하는 의지를 의미한다.

5.1.1.3 민주주의, 정의 및 법치 존중

헌법 민주주의 국가의 모든 정부 권력은 법률에 의해 합법적으로 승인되고 제한되어야 한다. 인공지능 시스템은 민주적 절차를 유지하고 육성하며 개인의 가치와 삶의 선택을 존중해야 한다. 인공지능 시스템은 민주적 절차, 인간의 숙고 혹은 민주적 투표 시스템을 훼손해서는 안 된다. 또한 인공지능 시스템은 법치주의가 확립한 기본 약속, 법률 및 규정을 훼손하지 않도록 보장하고 법 앞에 정당한 절차와 평등을 보장해야 한다.

5.1.1.4 평등, 차별 금지 및 연대 – 배제 위험에 처한 사람의 권리 포함

모든 인간의 도덕적 가치와 존엄성에 대한 동등한 존중을 보장해야 한다. 이것은 객관적 정당화에 근거하여 다른 상황 간의 구별을 용납하는 차별 금지를 뛰어넘는다. 인공지능 맥락에서 평등은 인공지능 시스템 운영이 부당하게 편향된 결과물을 생성할 수 없다는 것을 수반한다(예컨대, 인공지능 시스템을 훈련시키는 데 이용되는 데이터는 다른 표본 집단을 대표할 만큼 포괄적이어야 한다). 또한 취약한 사람과 집단, 여성, 장애인, 소수민족, 어린이, 소비자 또는 배제 위험에 처한 사람 같은 모든 사람을 보호한다.

5.1.1.5 시민의 권리

시민은 투표권, 선량한 행정이나 공공 문서에 대한 접근권, 행정부에 탄원할 권리 등 다양한 권리를 갖는다. 인공지능 시스템은 사회에 공공재와 공공 서비스를 제공하는 데 있어 정부의 규모와 효율성을 향상시킬 실질적인 잠재력을 제공할 수 있다. 동시에 인공지능 시스템은 시민의 권리에 부정적인 영향을 줄 수 있기 때문에 보호되어야 한다. 여기서 "시민의 권리"라는 용어를 사용하는 경우, 이는 인공지능 시스템의 영역에서 국제법상

의 권리를 가진 EU 역내 제3국 국민 및 불법인(또는 불법)의 권리를 부정하거나 무시할 수 없음을 의미한다.

5.1.2 인공지능 시스템 맥락에서 윤리 원칙

‘유럽을 위한 인공지능’보고서에서 인공지능에 대한 접근방식을 세 가지로 나누었다. 첫째, 공공 영역의 기술 및 민간 영역의 인공지능 기술 개발과 활용을 촉진시킨다. 둘째, 인공지능이 초래하는 사회 경제 변화에 대하여 대응한다. 셋째, 적절한 윤리 및 법체계의 마련과 보장이다. 아래에서 인공지능 윤리 및 법 관련 내용을 중심으로 정리하였다.

5.1.2.1 인공지능 시스템 맥락에서 윤리 원칙

공공, 민간 및 시민 단체는 인공지능 시스템의 윤리 프레임워크를 도출하는 데 기본권을 참조한다. 과학 및 신기술 윤리 강령 그룹(European Group of Science and New Technologies, 이하 "EGE")은 EU 조약 및 헌장에 명시된 기본 가치를 토대로 4 가지 기본 원칙을 제안하였다. EGE는 여러 그룹이 제시한 원칙을 인정함과 동시에 모든 원칙이 제시하는 목적을 명시하는 작업을 수행하고 있다. 이러한 윤리 원칙은 새롭고 구체적인 규제 수단에 단초를 제시할 수 있으며, 인공지능 시스템의 개발, 배포 및 사용에 대한 이론적 근거를 제시할 수 있다. 인공지능 시스템은 개인 및 공동 복지를 향상시켜야 한다. 인공지능 시스템의 신뢰를 구축하는 방식으로 개발, 배포 및 이용되는 것을 보장하기 위해 기본권에 대한 네 가지 윤리 원칙을 제시하였다.

4가지 원칙은 다음과 같다 :

(i) 인간자율성 존중
(ii) 피해 방지
(iii) 공정성
(iv) 설명가능성

4가지 원칙은 기존 법적 요구 사항에 이미 상당 부분 반영되어 있으므로 신뢰성있는 인공지능의 첫 번째 구성 요소인 합법적인 인공지능의 범위에 포함된다. 그러나 이와 같이 법적 의무 사항에 윤리 원칙이 반영되어 있지만 윤리 원칙을 준수하는 것은 기존 법을 공식적으로 준수하는 것 이상의 의미를 갖는다.

5.1.2.1.1 인간 자율성 존중의 원칙

EU의 기본권은 인간의 자유와 자율성 존중을 보장한다. 인공지능 시스템과 상호작용하는 인간은 인간 자신에 대한 완전하고 효과적인 자기결정을 유지할 수 있어야 하며 민주적 과정에 참여할 수 있어야 한다. 인공지능 시스템은 부당하게 인간을 종속시키고, 강제하

고, 조작하고, 조건화하거나, 무리지어서는 안 된다. 인공지능 시스템은 인간 인지 능력, 사회 문화 기술을 증강시키고, 보완하고, 강화시키도록 설계되어야 한다. 인간과 인공지능 시스템 간의 기능 할당은 인간 중심의 설계 원칙을 따라야 하며 인간 선택에 의미있는 기회를 제공해야 한다. 이는 인공지능 시스템의 작업 프로세스에 대한 인간의 감독권을 확보하는 것을 의미한다. 또한 인공지능 시스템은 작업 영역을 근본적으로 바꿀 수도 있다. 그것은 작업 환경에서 인간을 지원해야 하며, 의미있는 일자리 창출을 목표로 해야 한다.

5.1.2.1.2 피해 방지 원칙

인공지능 시스템은 인체에 해를 미치거나 악화시키거나 악영향을 주지 않아야 한다. 이것은 인간 정신 육체의 완전성과 인간 존엄성의 보호를 수반한다. 인공지능 시스템과 인공지능 시스템이 작동하는 환경은 위험 없고 안전해야 한다. 인공지능 시스템은 기술적으로 견고해야 하며 악의적 용도로 이용될 수 없다. 취약한 사람들에게 더 많은 주의를 기울여야 하고, 취약한 사람들에 대한 주의 내용이 인공지능 시스템의 개발, 배치 및 이용에 포함되어야 한다. 인공지능 시스템이 고용인과 피고용인, 기업과 소비자, 정부와 시민 등의 관계에서 권력 혹은 정보 비대칭으로 인해 악영향을 초래하거나 악화시킬 수 있는 상황에 주의해야 한다. 또한 피해를 방지하는 것은 자연 환경과 모든 살아있는 존재에 대한 고려를 수반한다.

5.1.2.1.3 공정성 원칙

인공지능 시스템의 개발, 배치 및 이용은 공정해야 한다. 공정성에 대한 다양한 해석이 있지만, 공정성에 실질적 차원과 절차적 차원이 모두 있다고 판단한다. 실질적 차원은 이익과 비용의 평등하고 공정한 분배를 보장하고 개인과 집단이 불공정, 차별 및 낙인으로부터 자유로울 수 있도록 보장하겠다는 약속을 의미한다. 불공정 행위를 피할 수 있다면 인공지능 시스템은 사회적 공정성을 향상시킬 수 있다. 교육, 재화, 서비스 및 기술에 대한 접근에서 동등한 기회를 조성해야 한다. 인공지능 시스템 이용이 선택의 자유에서 사기당하거나 부당하게 피해를 입은 사람에게 도달하면 안 된다. 또한 공정성은 인공지능 실무자가 수단과 방법 사이의 비례성 원칙을 존중하고 경쟁 목표와 이익을 조율시키는 방법을 주의깊게 고려해야 함을 의미한다. 절차적 차원의 공정성은 인공지능 시스템과 인공지능 시스템을 운영하는 인간의 결정에 대해 효과적으로 토론하고 교정할 수 있는 능력을 수반한다. 그렇게 하기 위해서는 의사결정의 실체를 구별할 수 있어야 하고 의사결정 과정을 설명할 수 있어야 한다.

5.1.2.1.4 설명가능성 원칙

설명가능성은 인공지능 시스템에 대한 이용자의 신뢰를 구축하고 유지 관리하는 데 중요하다. 프로세스가 투명해야 하며, 인공지능 시스템의 기능과 목적 및 의사 결정이 가능한

직접적이고 간접적으로 영향을 받는 사람들에게 설명할 수 있어야 함을 의미한다. 이런 정보가 없다면, 의사결정 이슈를 정당하게 논쟁할 수 없다. 어떤 모델이 특정 산출 혹은 결정 (그리고 그것에 기여한 입력 요소의 조합)을 생성한 이유에 대한 설명이 항상 가능한 것은 아니다. 이러한 경우를 '블랙박스' 알고리즘이라고 하며 특별한 주의를 기울여야 한다. 이런 상황에서 인공지능 시스템이 전체적으로 기본권을 존중하는 경우 다른 설명 가능성의 측정 방법 (예: 추적가능성, 감사가능성 및 시스템 기능에 대한 투명한 의사소통)이 필요할 수 있다. 설명가능성이 필요한 정도는 그 산출물이 잘못되었거나 부정확한 경우 결과의 내용과 심각도에 따라 달라진다.

5.1.3 신뢰할 수 있는 인공지능 구현을 위한 요건

5.1.3.1 자율성과 감독

1) 기본권. 시스템 개발에 앞서 인간의 권리와 자유를 존중하기 위해 기본권 영향 평가를 수행해야 한다.
2) 자율성. 인간의 자율성을 기반으로 인공지능 시스템을 사용할 수 있도록 설계되어야 하며, 인간이 시스템에 종속되지 않도록 해야 한다.
3) 감독. 인공지능 시스템의 적용 분야와 잠재적 위험에 따라 다양한 감독 체계가 필요하며, 정부는 해당 업무에 상응하는 감독을 수행할 수 있는 역량을 확보하여야 한다.

5.1.3.2 기술적 견고함과 안전성

1) 보안성. 해킹 등의 공격에서 보호받아야 하며, 이용자가 시스템을 악용할 수 있다는 점을 고려하여 이를 예방하고 완화해야 한다.
2) 안전성. 이용자, 자원 또는 환경에 피해가 되지 않도록 작업을 수행하고 의도하지 않은 결과와 시스템 작동 오류를 최소화해야 하며, 인공지능 제품·서비스 이용의 잠재적 위험을 평가하는 프로세스를 마련해야 한다.
3) 정확성. 인공지능 시스템은 데이터나 모델 기반으로 정확한 분류, 예측, 결정을 내릴 수 있는 능력을 필요로 한다.
4) 재현성. 인공지능 시스템이 동일한 조건 하에서 동일한 결과를 도출할 수 있어야 한다.

5.1.3.3 프라이버시와 데이터 거버넌스

1) 개인 정보와 데이터 보호. 프라이버시와 데이터(이용자가 제공한 데이터뿐만 아니라 인공지능 시스템과의 상호작용 과정에서 생성된 이용자 관련 모든 정보)의 보호는 인공지능 시스템의 전 생애주기 기간에 보장되어야 한다.
2) 데이터의 품질 및 무결성. 데이터가 부정확하거나 오류 및 편향적 성향을 포함하지

않도록 계획, 학습, 테스트 및 개발 등 각 단계에서 데이터와 프로세스를 테스트하고 문서화해야 한다.

3) 데이터 접근성. 개인 데이터를 처리하는 경우 데이터 접근을 관리하는 프로토콜을 배치해야 한다(정당한 자격을 갖춘 직원만 접근 허용).

5.1.3.4 투명성

1) 추적가능성. 투명성과 추적 가능성을 높이기 위해 데이터와 프로세스는 최상의 표준으로 문서화되어야 한다.
2) 설명가능성. 인공지능 시스템의 기술적 프로세스 관계자의 의사결정을 설명할 수 있어야 하며, 조직 의사결정 프로세스에 영향을 미치는 정도, 시스템 설계 선택 사항 및 시스템 배치 근거를 제공하여야 한다.
3) 의사소통. 인공지능 시스템은 이용자에게 스스로를 사람이라고 표현하지 않으며, 인공지능 시스템이라는 것을 사람이 인지할 수 있어야 한다.

5.1.3.5 다양성, 비차별성, 공평성

1) 비차별성. 개인 또는 단체별 특성 차이를 부당하게 이용하지 않고, 직·간접적 차별에 따른 편향성을 파악하여 부정적 영향을 최소화시킨다.

- 데이터는 항상 일정한 편견을 가지므로 편향성 수정, 차후에 수정될 수 있는 상향 식별은 인공지능 개발을 위해 중요하다.
- 인공지능 기술을 이용하여 내재적 편향성을 파악하고, 고유의 내재적인 편견에 대한 인식 개선 교육을 해야 한다.

2) 접근성 및 보편성 디자인. 인공지능 시스템은 성별, 나이, 국적, 장애 여부, 사회적 지위 등에 관계없이 모든 사용자가 서비스에 접근·이용할 수 있도록 설계되어야 한다.
3) 이해관계자 참여. 인공지능 시스템의 영향을 받을 가능성이 있는 이해관계자들의 의견을 반영하기 위한 방안을 마련해야 한다.

5.1.3.6 사회적 환경적 복지

1) 지속가능 환경친화적 인공지능. 모든 인공지능 시스템 공급 과정에서 발생하는 자원 사용 및 에너지 소비를 비판적으로 검토하여 환경적으로 유해한 결과를 방지해야 한다.
2) 사회적 영향. 인공지능 시스템은 사회적 기술을 향상시키는 동시에 악화하는 결과를 가져올 수 있으므로 인공지능 시스템의 사회적 영향을 고려하고 주의 깊은 모니터링을 해야 한다.

3) 사회와 민주주의. 정치적 의사결정, 선거 상황을 포함한 민주적 과정과 관련된 상황에서 인공지능 시스템의 활용은 신중히 고려되어야 한다.

5.1.3.7 책임성

1) 금전적 보상(무과실 보험). 과실 책임, 금전적 보상이 없는 화해 등 인공지능 거버넌스에 책임 메커니즘을 포함시켜야 한다.
2) 감사. 알고리즘, 데이터 및 설계 프로세스 평가를 위한 인공지능 시스템에 대한 독립적 감사를 필요로 한다.
3) 부정적 영향의 최소화와 문서화. 인공지능 시스템의 부정적인 영향을 식별하고, 평가한다. 식별과 평가 내용을 문서화한다. 인공지능 시스템의 부정적인 영향을 최소화하기 위하여 인공지능 시스템 개발, 배포 전 및 배포 중에 영향 평가 문서를 활용한다.
4) 상충 관계. 제시된 요건들의 충돌이 있는 경우 최신 기술 수준 내에서 합리적으로 상충 관계를 설명할 수 있어야 한다.
5) 구제. 부작용 발생 시, 적절한 배(보)상을 보장할 수 있는 메커니즘을 마련하여 신뢰를 확보해야 한다.

5.2. IEEE의 윤리적으로 조율된 설계(EAD)

5.2.1 자율적 지능 시스템에 반영되어야 할 윤리적 고려사항

인공지능이 인간의 삶과 비즈니스에 어떤 영향을 주는 중요한 논의 주제가 되었다. 국제전기전자학회(IEEE)는 인공지능 기술 개발에 있어서 인간의 복리를 우선하기 위한 3가지 판단 도구를 표준화하는 작업을 진행 중이다. 일상에 사용되는 인공지능 및 자동화 시스템이 윤리적인가를 판단하는 표준을 제정하여 인간의 복리와 안전을 보장하겠다는 의도이다. 다시 말해서 점차 자동화 및 인텔리전트 시스템이 일상 생활에서 더 중요한 역할을 수행하게 되므로 자동화된 시스템에 적용된 기술들은 윤리적 고려사항이 반영되어야 하고 이를 측정할 표준도 요구된다.
이러한 배경에서 국제전기전자학회(IEEE)산하의 자율적 지능 시스템의 윤리적 설계를 위한 글로벌 이니셔티브는 지능형 제품, 서비스 개발을 위한 가이드라인을 개발하였다. 2017 년에 발표된 ‘윤리적으로 조정된 설계’는 글로벌 이니셔티브 산하 13개 위원회 의견을 반영하여 작성되었다. ‘윤리적으로 조정된 설계’ 문서에 포함되어 있는 이슈, 권고안들은 자율적 시스템이 인간 사회의 안전, 복지, 이익을 증진시킬 수 있도록 윤리적 고려사항을 구체화하여 인공지능 기술개발의 방향을 유도하는 기준으로 기능한다고 할 수 있다. 동 가이드라인이 표명하는 일반 원칙은 다음과 같다.:

1) 인권: 자율적 지능 시스템의 윤리적 설계, 개발, 실행은 국제적으로 인정 되는 인권을 인정해야한다
2) 복리: 자율적 지능 시스템의 윤리적 설계, 개발, 실행은 인간 복리의 측정을 우선시해

야 한다.
3) 책무성: 자율적 지능 시스템의 윤리적 설계와 조작자는 책임과 책무성이 요구된다.
4) 투명성: 자율적 지능 시스템의 윤리적 설계, 개발, 실행은 투명해야 한다.
5) 오남용에 대한 인식: 자율적 지능 시스템의 윤리적 설계, 개발, 실행은 오남용을 최소화해야 한다.

5.2.2 '윤리적으로 조율된 설계(EAD)'와 법적 규제 권고

자율적 지능 시스템의 개발, 활용과 관련하여 많은 논의가 오가고 있지만 정작 법적으로 어떻게 다루어야 하는지는 아직 결정되지 않았다. 자율적 지능 시스템의 법적 규제와 관련 하여 '윤리적으로 조율된 설계(EAD)'는 다음과 같은 4가지 영역에 걸쳐서 초점과 함께 권고 안을 제시하고 있다. 이하의 질문은 공개적 토론을 유도하기 위한 것이며 법률가 커뮤니티의 참여 등 공개적 의견 수렴을 거쳐서 점차 구체화될 것으로 전망된다.:

1) 자율적 지능 시스템의 법적 지위
2) 정부가 자율적 지능시스템을 이용하는 경우: 투명성과 개인의 권리·자율적 지능 시스템에 의해 초래된 해악의 법적 책임
3) 자율적 지능 시스템의 투명성, 책무성, 확증가능성

또한, 최근 2018년 7월 IEEE와 IEEE 표준 협회(IEEE Standards Association)는 자율 및 지능형 시스템의 윤리를 위해 공개 커뮤니티(OCEANIS, Open Community for Ethics in Autonomous and Intelligent Systems)를 9개 표준화 기구들(SDO)과 함께 시작하였다. 이는 정보통신 기술 분야에서 개발되는 자율 및 지능형 시스템이 윤리적으로 조화될 수 있도록 표준을 개발하기 위한 협력을 촉진하는 글로벌 포럼이다. OECIANIS는 인공지능 기술과 관련한 비즈니스 및 정책결정을 지원하는 표준을 개발하기 위한 모든 관련 국제 조직 간의 광범위한 협력을 도모한다. OECIANIS는 자율 및 지능형 시스템의 개발 및 보급과정에서 나타나는 복잡한 윤리적 문제는 비공개 협약을 통해서가 아니라 공동의 협의를 통해서만 해결될 수 있다고 전제하고 있다. 이는 IEEE의 자율 및 지능형 시스템 윤리 글로벌 이니셔티브 (Global Initiative on Ethics of Autonomous and Intelligent Systems)의 방향과도 일치한다.

5.2.3 인공지능에 적용할 수 있는 윤리적 고려사항

5.2.3.1 일반 원칙

일반 원칙은 모든 유형의 인공지능/자율시스템에 응용할 수 있는 가장 높은 수준의 윤리적 고려사항은 다음과 같다 :

1) 인권의 가장 높은 이상형을 구체화한다.

2) 인간성과 자연 환경의 최대 이익을 우선시한다.
3) 인공지능/자율시스템이 사회-기술 시스템으로 진화함에 따라 위험과 부정적 영향을 완화시킨다.

5.2.3.2 인공지능과 자율시스템 윤리 이슈

윤리 이슈는 다음과 같다:

1) 인공지능과 자율시스템이 인권을 침해하지 않게 하는 방안은?
2) 인공지능과 자율시스템에게 책임성을 부여할 방안은?
3) 인공지능과 자율시스템의 투명성을 높이는 방안은?
4) 인공지능과 자율시스템의 위험을 최소화시키고 이익을 극대화시키는 방안은?

5.2.4 가치를 자율지능 시스템에 내장시키기

인공지능 서비스 제공자가 사회에 이익을 제공하는 인공지능과 지율시스템을 성공적으로 개발하려면 연관성있는 인간 규범 혹은 가치를 인공지능과 자율시스템에 내장시키는 것에 대한 이해와 관련 능력을 보유해야 한다. 이를 3개 영역과 영역별 이슈로 나누어 볼 수 있다:

1) 자율지능 시스템에 적용하기 위한 규범과 가치의 확인
2) 자율지능 시스템에 대한 규범과 가치를 내장시키는 방안은?
3) 인간과 자율지능 시스템간 규범과 가치의 조화에 대한 평가

자율지능 시스템에 적용하기 위한 규범과 가치의 확인에 관한 세부 이슈는 다음과 같다:

1) 자율지능 시스템에 내장되는 가치는 보편적인 것이 아니라 이용자가 속한 공동체 및 작업 대상에 따라 달라져야 한다.
2) 도덕 과잉이 있다. 도덕과잉은 자율지능 시스템에 여러 유형의 상충하는 윤리가 구현될 수 있는 것을 의미한다.
3) 자율지능 시스템은 특정 집단에 대한 편견을 가진 자료나 알고리즘을 구현할 수 있다.

자율지능 시스템에 대한 규범과 가치를 내장시키는 방안은? 에 관련 세부 이슈는 다음과 같다:

1) 특정 가치나 규범을 확인한 후 이것을 어떤 방식으로 컴퓨터 시스템에 설계할 것인가는 아직 분명하지 않다.

인간과 자율지능 시스템간 규범과 가치의 조화에 대한 평가에 관한 세부 이슈는 다음과 같다:

1) 자율지능 시스템에 구현되는 규범은 관련 커뮤니티의 규범과 양립 가능해야 한다.
2) 인간과 자율지능 시스템간 신뢰에 대한 정확한 수준을 마련해야 한다.
3) 자율지능 시스템의 가치 조화에 대한 제3자 평가가 필요하다.

5.2.5 자율 시스템 관련 윤리 연구와 자율 시스템의 설계를 가이드하는 방법론

인공지능/자율시스템 관련 조직은 인간 복지, 권한이양, 자유가 인공지능/자율시스템 개발의 핵심임을 알아야 한다. 윤리적 건전성을 중심으로 하는 접근방법이 인공지능의 경제적/사회적 적정성을 가능하게 한다. 자율시스템 관련 윤리 연구와 자율시스템의 설계를 가이드하는 방법론 관련 세부 이슈는 다음과 같다:

1) 학제간 교육과 연구
2) 비즈니스 관행과 인공지능
3) 투명성의 부제

학제간 교육과 연구의 세부 이슈는 다음과 같다:

1) 현재 윤리가 지율시스템 관련 교육 프로그램의 한 부분이 아니다.
2) 인공지능/자율시스템의 고유 특징을 이해할 수 있는 학제적/다문화 교육 모델이 필요하다.
3) 인공지능/자율시스템 설계 시 구현되는 문화의 다양한 가치를 구분할 필요가 있다.

비즈니스 관행과 인공지능의 세부 이슈는 다음과 같다:

1) 산업계에 가치 기반의 윤리 문화 및 관행이 아직 수립되어 있지 않다.
2) 가치에 대한 인식 고양을 위한 리더쉽이 부재하다.
3) 윤리적 관심에 대한 자원이 부재하다.
4) 기술 공동체의 주도권 및 책임감이 부재하다.
5) 최고의 인공지능/자율시스템을 설계하기 위한 다양한 이해관계자의 참여가 필요하다.

투명성의 부재 관련 세부 이슈는 다음과 같다:

1) 부실한 절차 관리가 윤리적 설계를 방해한다.
2) 알고리즘에 대한 감독이 부제하고, 일관성이 부족하다.

3) 알고리즘에 대한 독립 리뷰 조직이 부재하다.
4) 시스템 설계 시 블랙박스 도구의 사용으로 인해 문제가 발생한다.

5.2.6 인공지능의 안전

미래의 고성능 인공지능 시스템은 세계적 수준에서 농업 혹은 산업 혁명에 대변화의 영향을 줄 것이다. 인공지능/자율시스템의 안전과 혜택에 대한 이슈가 발생한다. 인공지능/자율시스템의 안전과 혜택에 대한 이슈는 다음과 같다 :

1) 기술 문제.
2) 일반 원칙.

기술 문제에 대한 세부 이슈는 다음과 같다:

1) 자율시스템이 다양한 상황을 학습된 경험으로 처리할 수 있는 증대된 자율성으로 복잡한 기능을 최적화하여 수행할 수 있게 됨으로써 예상치 못한 혹은 기대되지 않은 행동으로 인해 위험한 상황이 발생할 수 있다.
2) 미래의 강력한 인공지능에 안전을 장착하기 쉽지 않다.

일반 원칙에 대한 세부 이슈는 다음과 같다:

1) 개발자들은 자율화되고 효율적인 인공지능 시스템을 개발하거나 운영할 시 복잡한 윤리 및 기술 안전 이슈들에 직면할 것이다.
2) 미래 인공지능 시스템은 농업 및 산업 혁명에 준하는 영향력을 사회에 행사할 것이다.

5.2.7 개인 데이터와 개별 접근 통제

개인 정보에 관한 핵심 윤리 딜렘마는 데이터 비대칭성이다. 개인 데이터 비대칭성을 해결하려면 이용자가 자신의 개인정보를 통제할 수 있는 도구를 보유해야 한다. 개인 데이터와 개별 접근 통제 관련 세부 이슈는 개인 정보에 대한 정보, 개인 정보 접근 및 동의, 개인 자료 관리이다.

개인 정보에 대한 정보 관련 세부 이슈는 다음과 같다:

1) 알고리즘 구현 시 무엇을 개인 정보라고 정의할 것인가?
2) 개인 식별 가능 정보에 대한 정의 및 범위는 무엇인가?
3) 개인 정보에 대한 통제에서 통제의 정의는 무엇인가?

개인 정보에 대한 세부 이슈 관련 권고 사항은 다음과 같다:

1) 개인들은 신뢰형 자기동일성 확증 리소스를 확인하여 자신의 자기동일성을 확인, 실증, 제공할 수 있어야 한다.
2) 개인 식별 정보(PII, Personal Identifiable Information)는 개인의 유일한 물리적, 디지털, 혹은 가상 자기동일성에 상관되어 있는 개인에 합리적으로 연결된 모든 데이터로 정의된다.
3) 개인 데이터는 이용자의 인식 혹은 이용자의 통제를 벗어나 이용자의 개인 데이터에 접근하는 행위자가 아닌 이용자 본인의 관점에서 관리되어야 한다.

개인 데이터 접근 및 동의에 관한 세부 이슈는 다음과 같다:

1) 개인 존중을 위해 자료에 대한 접근을 어떻게 재정의해야 하나?
2) 개인 데이터 이용에 대한 동의를 어떻게 재정의 해야 하나?
3) 사소한 개인 데이터이지만 그 개인 데이터가 공유될 때 개인이 공유를 원치 않는 개인 정보를 추론할 수 있다.
4) 데이터 처리자가 자료 접근 및 수집의 부정적 혹은 긍정적 결과를 개인에게 완전하게 알릴 수 있는가?

개인 데이터 관리에 대한 세부 이슈는 다음과 같다:

1) 이용자가 개인화된 인공지능 혹은 알고리즘 후견인을 가질 수 있는가?

5.2.8 인도주의 이슈

인공지능 시스템은 신뢰와 안전 환경에서 채택되기 때문에 리소스의 투명성, 명확성, 가용성을 증가시키는 노력을 필요로 한다.

5.2.8.1 책임과 규칙에 맞는 이용

개인 데이터 이용에 관한 책임과 개인 데이터의 규칙에 맞는 이용에 대한 세부 이슈는 다음과 같다:

1) 개인 데이터에 대한 접근 및 이해 부족

개인 데이터 접근에 대한 책임과 개인 데이터의 규칙에 맞는 이용에 대한 세부 이슈에 대한 권고 사항은 다음과 같다:

1) 무해 행동 지침이 비상 상황과 이해충돌 상황에 적용되어야 한다.

5.3 ACM의 윤리 강령과 전문가 행동 강령

5.3.1 주요 활동

ACM은 컴퓨터 학회로 2017년에 '알고리즘 투명성과 책무성에 대한 성명'을 발표하였다. 이 성명에는 인공지능 시스템의 소유자, 설계자가 고려해야할 사항들이 담겨있다. 2018년에는 컴퓨터 기술이 사회 전반에 심대한 영향력을 발휘하고 있으므로 엔지니어의 책임이 변화되었다고 밝히면서 강화된 윤리 강령 및 전문가 행동 강령(ACM Code of Ethics and Professional Conduct)을 공개하였다. 성명과 행동 강령은 자율적 가이드라인이어서 법적 구속력은 없지만, 인공지능의 윤리와 긴밀한 관계에 있는 알고리즘 투명성과 책무성 관련 가이드라인 역할을 할 것으로 보인다.

5.3.2'알고리즘 투명성과 책무성에 대한 성명'

1) 인식: (인공지능 알고리즘이 적용된)분석 시스템의 소유자, 설계자, 이용자 및 기타 이해관계자들은 설계, 실행, 이용 과정에서 발생할 수 있는 편향의 가능성들을 인식하여야 한다. 편향은 개인과 사회에게 해악을 초래할 수 있기 때문이다.
2) 접근과 수정: 조정자들은 알고리즘 기반 의사결정에 의해 악영향을 받는 개인 및 집단이 의문을 제기하고 바로 잡을 수 있는 메커니즘을 채택하도록 장려해야 한다.
3) 책무성: 기관들은 자신들이 이용하는 알고리즘 기반 의사결정에 대해 책임을 져야 한다. 알고리즘이 결과를 산출하는 방식에 대해 자세히 설명할 수 없는 경우에도 책임을 져야 한다.
4) 설명가능성: 알고리즘 기반 의사결정을 이용하는 시스템 및 기관들은 알고리즘이 따르는 프로시저 및 수행된 특정 결정들에 대한 설명을 작성한다. 이러한 설명은 공공 정책의 맥락에서 매우 중요하다.
5) 데이터 출처 파악: 알고리즘 작성자는 (인공지능 알고리즘에 공급된)훈련용 데이터를 수집한 방법에 대한 기술을 유지해야 하고, 인간 또는 알고리즘에 의한 데이터 수집 절차에 의해 유도될 수 있는 잠재적인 편향들을 조사해야 한다. 공개적으로 데이터를 조사한다면 수정할 수 있는 기회가 최대치로 보장되어야 한다. 그러나 개인 정보 및 영업 비밀 보호에 대한 우려, 분석 방식의 노출 우려 또는 악의적 행위자에 의해 시스템이 조작될 우려가 있으므로, 자격을 갖추고 인가된 개인에게만 조사를 위한 접근을 제한한다.
6) 감사가능성: (인공지능)모형, 알고리즘, 데이터 및 의사결정이 위험을 발생시킬 가능성에 대비하여 감사받을 수 있도록 기록되어야 한다.
7) 유효성 검사와 테스트: 연구개발 주체들은(인공지능)모형의 유효성을 검사하기 위하여 엄격한 방법을 이용해야 하고 검사 방법과 검사 결과를 문서화해야 한다. 특히,(인공지능) 모형이 차별적 해악을 입히는 지 여부를 평가하고 결정하기 위한 검사를 정기적으로

수행해야 한다. 검사결과를 공개하도록 권장한다.

5.3.3 윤리 강령

전미컴퓨터학회(ACM)가 2018년에 개정한 윤리 강령(Code of Ethics)은 자율적인 행태 준칙이기 때문에 강제성은 없다. 윤리강령1조는 기본 윤리 원칙, 2조는 전문가의 책임에 대한 설명을 담고 있다. 3조는 리더십 역할을 수행하는 전문가 관련 내용이다. 이 윤리 강령은 컴퓨터 전문가들을 대상으로 발표되었고, 실무자들이 연구개발 과정에서 스스로 윤리적 결정을 내리도록 일정한 행동 준칙을 제시하고 있다. 이하는 새로운 윤리 강령 1조 ~ 2조 내용 중에 인공지능 윤리와 관련있는 조항을 정리하였다.

5.3.3.1 피해의 회피

1.1 컴퓨터 전문가들은 컴퓨터 기술로 인하여 발생하는 피해를 최소화해야만 한다. 피해는 건강, 안전, 신체의 안전, 개인 정보(프라이버시) 침해에 관련된 위협을 의미한다.

1.2 '피해'는 정당하지 않은 물리적·정신적 손상, 정당하지 않은 정보의 파괴, 정보의 공개, 재산, 명성에 대한 정당하지 않은 손상 등이다. 실수 또는 의도하지 않았지만 피해가 발생한 경우, 그 행위의 책임자는 피해를 최대한 복구시키기 위해 노력해야 한다. 피해로 이어질 수 있는 시스템 위험의 징후가 발견되면 즉시 알려야한다. 만일 결정권자가 보고를 받고서도 아무런 조치를 취하지 않는다면 사법당국에 고발해야 한다.

5.3.3.2 투명성과 공정성

1.3 컴퓨터 전문가들은 시스템 역량과 한계, 발생가능한 문제들을 투명하게 완전히 공개해야 한다. 이에 대하여 고의적으로 허위 주장을 하거나 오해를 불러일으키거나, 데이터를 조작하거나 위조해서는 안 된다.

5.3.3.3 차별에 대한 주의

1.4 새로운 기술의 도입과 활용은 기존에 없던 차별을 야기할 수 있다. 그러므로 새로운 사항을 결정하고 시도한다면 사후에도 관심을 유지하고 현상을 관찰해야 한다.

5.3.3.4 개인 정보의 존중

1.6 각종 기술의 발전으로 인하여 개인 정보는 위협받고 있다. 최신 기술은 개인의 정보를 빠르게 수집하고, 모니터링하고, 교환한다. 이러한 데이터의 흐름 속에서 당사자도 모르게 개인 정보의 침해가 일어날 수 있다. 그러므로 개인 정보가 어떻게 수집되고 활용되는 가를 이해하고, 정당한 권리를 가진 자들에게 그 과정을 설명해 주어야 한다. 개인

정보의 이용은 법 규제의 테두리를 벗어나서는 안 된다. 수집하고 활용하는 개인정보가 식별 가능 정보로 변할 수 있는지 여부, 유통되는 과정 중에 유출될 수 있는 지 여부를 주의 깊게 살펴야 한다. 데이터의 변질이나 불법적 접근, 실수로 인한 유출의 가능성도 최대한 줄여야 한다.

5.3.3.5 대중 인식

2.7 대중들이 컴퓨터 관련 기술들을 보다 더 잘 이해하고 인지할 수 있도록 돕고, 알리고, 권장해야 한다. 특히 컴퓨터 시스템의 영향력과 한계, 취약점, 기회를 쉽고 친절하고 알기 쉽게 알려줄 수 있어야 한다.

5.3.3.6 안전

2.9 컴퓨터 시스템을 설계하거나 도입할 때 원래 의도한 대로 기능을 발휘하도록 하고 사고나 악의에 의한 남용, 조작, 마비 상황에 미리 대비하고 최대한 차단해야 한다. 새로운 시스템의 도입으로 인하여 취약점들이 생겨나고 새로운 공격 가능성도 증가한다. 그러므로 이에 대비하기 위한 장치 및 정책을 마련하는 것도 컴퓨터 전문가들의 책임이다. 여기에는 취약점 모니터링, 패치, 취약점에 대한 보고 등이 포함된다.

5.4 Internet Society의 인공지능과 기계학습

Internet Society는 인터넷의 발전을 위해 설립된 국제 비영리 기구로서 모든 인터넷 관련 기구들 가운데 최상위에 있다. 인터넷 기술 개발이나 운용 관리상의 문제들을 총괄 관리한다. 인터넷 소사이어티는 2017년 '인공지능 및 기계학습에 대한 정책 보고서'를 발간하였다. 보고서의 내용 중에 '인공지능 설계, 이용의 윤리적 고려'와 '책임있는 이용'은 인공지능 윤리, 규범 정책과 관련성이 있어, 그 내용을 정리한다.

5.4.1 인공지능 시스템의 설계와 배포에서 윤리적 고려사항

5.4.1.1 원칙

인공지능 시스템 설계자는 기술에 '이용자 중심의 접근법'을 응용할 필요가 있다. 인공지능 시스템 설계자는 인터넷과 인터넷 이용자들이 보안 위험에 노출되지 않도록 '집단적 책임'을 인공지능 시스템에 포함시킨다.

5.4.1.2 권고 사항

1) 윤리 표준의 채택: 윤리 고려를 위한 원칙 및 표준의 준수가 필요하며, 연구자들 및 산업계가 나아갈 발전 방향이 제시되어야 한다.

2) 혁신 정책에서 윤리 고려의 촉진 자금 조달의 선결 조건으로 윤리적 표준을 준수할 것을 혁신 정책에 포함시켜야 한다.

5.4.2 책임 배포

5.4.2.1 원칙

인공지능 행위자는 자율적으로 행동하고 인간의 지시 없이도 시간이 흐르면서 인공지능 행위자의 행동을 조정할 수 있는 능력이 요구된다. 인공지능이 배포되기 이전에 이런 역량이 안전한지 여부를 점검하고 지속적인 모니터링을 해야 한다.

5.4.2.2 권고 사항

1) 인간의 통제. 인간이 자율 시스템의 작동을 중단시키거나 자율 시스템을 종료시킬 수 있어야 한다(“작동 종료 스위치”). 특히 인간의 생명과 안전에 위험을 미치는 인공지능 시스템 설계에 새로운 자율 의사결정 방식을 도입하는 경우에, 인간의 통제 능력을 인공지능 시스템에 통합시켜야 한다.
2) 안전 최우선. 모든 자율 시스템은 배포에 앞서 광범위한 테스트를 받아야 한다. 인공지능 행위자가 디지털 환경 혹은 물리적 환경과 안전하게 상호작용하고 목적에 부합하게 작동하는 지를 확인해야 한다. 자율 시스템은 작동 중에 모니터링되고 필요하면 업데이트 또는 수정되어야 한다.
3) 개인정보의 보호. 인공지능 시스템은 데이터 책임성이 요구된다. 인공지능 시스템은 필요한 데이터만 이용하고 데이터가 더 이상 필요하지 않다면 삭제해야 한다 (“데이터 최소화"). 인공지능 시스템은 데이터의 전송 중 및 유휴 중에 데이터를 암호화해야 하며, 인가된 사람만 접근할 수 있도록 통제되어야 한다(“데이터 접근 통제”). 인공지능 시스템은 개인 정보 법 및 모범 사례에 따라 데이터를 수집, 사용, 공유, 저장해야 한다.
4) 데이터 공급의 신중성. 인공지능 시스템에 제공하는 지침과 데이터에 대한 신중한 고려가 필요한다. 인공지능 시스템은 편향되거나, 부정확하거나, 불완전하거나, 오해의 소지가 있는 데이터로 훈련되어서는 안 된다.
5) 인공지능 시스템의 안전성. 인터넷에 연결된 인공지능 시스템은 안전하게 유지되어야 한다. 인공지능 시스템과 인터넷 모두를 보호하기 위해 안전은 중요하다. 맬웨어에 감염된 인공지능 시스템은 차세대 봇넷(Botnet)이 될 수 있다. 인공지능 시스템의 기기, 시스템 및 네트워크 보안에는 상향형 표준이 적용되어야 한다.
6) 책임 배포. 선의로 행동하는 보안 연구자는 기소 또는 법적 조치에 대한 우려 없이 책임있게 인공지능 시스템의 보안을 검사할 수 있어야 한다. 보안 취약점이나 기타 설계 결함을 발견하면 그 문제를 해결가능한 최상의 위치에 있는 사람이 책임있는 배포를 해야 한다.

5.5 아실로마 인공지능 23 원칙

옥스퍼드 대학의 미래 삶 연구소(FLI)는 2017년 '유익한 인공지능 콘퍼런스'(Beneficial AI Conference)를 개최하였다. 컨퍼런스에 참여한 경제, 법률, 윤리 및 철학 분야 전문가들이 인공지능 윤리를 중심으로 토론하였다. 이 토론을 통해 아실로마 인공지능 23 원칙을 도출하였다. 아실로마 인공지능 23가지 원칙은 연구 이슈, 윤리 및 가치, 장기 이슈 3부분으로 구성되어 있다. 3부분 가운데 윤리와 가치와 직결된 항목들은 13가지이다.

5.5.1 윤리 및 가치

6)안전. 인공지능시스템은 작동 수명 전반에 걸쳐 안전해야 하며, 인공지능 시스템이 적용 가능하고 실현가능할 경우 그 안전을 검증할 수 있어야 한다.
7)장애 투명성. 인공지능 시스템이 손상을 일으킬 경우 그 이유를 확인할 수 있어야한다.
8)사법적 투명성. 사법 제도 결정에 인공지능 자율 시스템이 사용된다면, 권위 있는 인권기구가 감사할 경우 만족스런 설명을 제공할 수 있어야 한다.
9)책임. 고성능 인공지능 시스템의 설계자는 인공지능의 사용, 오용 및 행동의 도덕적 영향에 관한 이해관계자이며, 이에 따라 그 영향을 형성하는 책임과 기회를 가진다.
10)가치 정렬. 고성능 자율 인공지능 시스템은 작동하는 동안 고성능 자율 인공지능 시스템의 목표와 행동이 인간의 가치와 일치하도록 설계되어야 한다.
11)인간의 가치. 인공지능시스템은 인간의 존엄성, 권리, 자유 및 문화적 다양성의 이상에 적합하도록 설계되어 운용되어야 한다.
12)개인정보 보호. 인공지능 시스템에서 이용되는 데이터를 분석하는 능력 및 활용하는 능력이 있다는 전제 하에, 인간은 그 자신들이 생산한 데이터에 대한 접근, 데이터 관리 및 데이터 통제할 수 있는 권리를 가져야 한다.
13)자유와 개인 정보. 개인정보에 관한 인공지능의 쓰임이 인간의 실제 또는 인지된 자유를 부당하게 축소할 수 없다.
14)공동 이익. 인공지능 기술은 최대한 많은 사람에게 혜택을 주고 힘을 실어주어야 한다.
15)공동 번영. 인공지능에 의해 이루어진 경제 번영은 인류의 모든 혜택을 위해 널리 공유되어야 한다.
16)인간의 통제력. 인간이 선택한 목표를 달성하기 위해 인간은 의사결정을 인공지능 시스템에 위임하는 방법 및 여부를 선택해야 한다.
17)비파괴. 고성능 인공지능 시스템의 통제로 주어진 능력은 건강한 사회가 지향하는 사회적 및 시정 과정을 전복하는 것이 아니라 그 과정을 존중하고 개선해야 한다.
18)인공지능 무기 경쟁. 치명적인 인공지능 무기의 군비 경쟁을 피해야 한다.

5.6 구글의 인공지능 원칙

구글은 AI 연구 및 활용이 사회에 중대한 영향을 줄 것으로 예측하고, 인공지능 관련 비

즈니스 의사결정에 영향을 주는 구체적인 기준 7가지를 발표하였다. 7가지 원칙 중에 윤리 영역과 관련있는 1~5 원칙을 정리하였다. 그리고 인공지능 기술이 활용되는 영역을 제시한 것을 정리하였다.

5.6.1 인공지능 7 원칙

1) 사회적 유익성

인공지능은 의료, 보안, 에너지, 운송, 제조 및 엔터테인먼트 등 분야에서 혁신적 영향을 미칠 것이므로 광범위한 사회적, 경제적 요인을 고려하여 전반적 이익이 예상되는 위험과 단점을 상회한다고 판단될 때 활용할 것이다. 국가의 문화, 사회규범, 법규범을 존중하면서 고품질의 정보를 쉽게 이용할 수 있도록 노력한다.

2) 불공정한 편향의 데이터 생성. 강화의 금지

인공지능 알고리즘과 데이터 집합은 불공정한 편견을 반영, 강화 또는 감소시킬 수 있다. 공평하고 공정한 판단은 단순하지 않고 문화와 사회에 따라 다르므로 구글은 인종, 성별, 국적, 소득, 성적 취향, 능력 및 정치적. 종교적 신념과 같은 민감한 특성에 대한 부당한 영향력을 주지 않기 위해 노력할 예정이다.

3) 강력한 안전 및 보안의 실행

인공지능이 유해한 위험을 결과를 야기하지 않도록 강력한 안전 및 보안 방법을 개발하여 실행할 예정이다. 인공지능 시스템을 적절하게 신중하게 설계하고 인공지능 안전 연구의 모범 사례에 따라 개발할 것이다. 제약 조건이 있는 환경에서 인공지능 기술을 테스트하고, 배포 후에도 작업을 모니터링할 예정이다.

4) 책임을 지는 인공지능 시스템

인공지능에 대한 피드백, 관련 설명 및 불만 제기를 위한 적절한 기회를 제공하는 시스템을 설계할 예정이다. 인공지능 기술은 적절한 인간의 지시와 통제를 받도록 할 것이다.

5) 개인정보 보호 원칙을 인공지능 설계에 통합하기

개인 정보 보호 원칙을 인공지능 기술 개발 및 사용에 반영할 예정이다. 이용자에게 통지하여 동의 기회를 제공하고, 프라이버시 보호 장치가 있는 아키텍처를 장려하며, 데이터 이용에 대한 적절한 투명성과 제어를 제공한다.

5.6.2 인공지능 기술 활용이 금지되는 분야

1) 전반적으로 해악을 주거나 일으킬 가능성이 있는 기술. (중대한 위해 위험이 있다면 이익이 중대한 위험보다 우위 있다고 믿는 분야에서만 진행하고 이 경우에도 적절한 안전 제약을 부과)
2) 인명 피해를 야기하거나 직접적인 손상을 가하는 목적의 무기 또는 기타 기술.
3) 국제 규범을 위반하는 감시 활동(Surveillance)을 통해서 얻어진 정보를 사용하는 기술·국제법 및 인권 원칙에 위배되는 기술.

6. 요약

윤리는 인간의 생활에 있어 바람직한 상태란 무엇이며, 좋은 것과 나쁜 것의 기준은 무엇이고, 행위의 법칙은 어떻게 정립되는가와, 노력할 만한 것은 무엇이며, 생활의 의미라는 것은 무엇인가 등을 밝히는 것으로 정의된다. 이 정의는 인공지능과 인공지능 시스템이라는 인공물이 인간의 생황에서 있어 바람직한 상태, 인공지능 시스템과 인간과의 좋은 관계 설정, 인공지능 시스템 설계와 배포에서 노력할 만한 것은 무엇인지 등을 밝히는 것으로 해석할 수 있다.
이 기술 보고서는 인공지능 서비스 제공자에게 인공지능 관련 윤리에 대한 전반적인 가이드라인 동향을 제시하고, 인공지능 서비스 제공자가 인공지능 시스템을 디자인하고 개발하는 경우에 이 보고서를 참조하여 윤리적 고려사항을 우선시할 수 있도록 하는 데 있다. 아래의 표는 주요 인공지능 윤리 가이드라인의 주요 내용과 특성을 정리한 것이다.

<표 1> 인공지능 윤리 가이드라인의 주요 내용과 특징

	주요 내용	특징
1.EU의 믿을만한 인공지능을 위한 윤리 가이드라인	인공지능 시스템의 신뢰를 구축하는 방식으로 개발, 배포 및 이용되는 것을 보장하기 위해 기본권에 대한 4 가지 윤리 원칙(인간자율성 존중, 피해방지, 공정성, 설명가능성)을 제시하였다. 또한 신뢰할 수 있는 인공지능 구현을 위한 요건을 제시하였다. 자율성과 감독, 기술적 견고함과 안전성, 프라이버시와 데이터 거버넌스, 투명성, 다양성/비차별성/공평성, 사회적 환경적 복지, 책임성을 요건으로 제시하였다. 4가지 원칙은 기존 법적 요구 사항에 이미 상당 부분 반영되어 있으므로 윤리 원칙을 준수하는 것은 기존 법을 공식적으로 준수하는 것 이상의 의미를 부여하는 인공지능 윤리 가이드라인이다.	자율적인 행태 준칙으로 제시하였으나 EU의 제도 특성상 강제성 있는 법률 제정을 할 것으로 예측된다.
2. IEEE의 윤리적으로 조율된 설계	지능형 제품, 서비스 개발 관련 윤리적 설계 가이드라인을 제시하기 위해 일반 원칙 5가지(인권,	IEEE와 IEEE 표준협회,

(EAD)	복리, 책무성, 투명성, 오남용에 대한 인식)를 정하고. 자동화된 시스템에 응용된 기술들에 윤리적 고려사항을 반영하고 이를 측정할 표준을 제정하는데 초점을 두고 있다. 특히, 윤리적 가치를 자율지능 시스템에 내장시키기 부분은 인공지능 시스템 개발과 연관성이 매우 높다.	OCEANIS, SDO 는 협업을 통해 인공지능 관련 국제 법 제정을 의도하고 있다.
3.ACM의 윤리강령과 전문가 행동 강령	강화된 윤리 강령 및 전문가 행동 강령에서 인공지능 윤리와 연관있는 알고리즘 투명성과 책무성 관련 가이드라인을 제시하였다. 알고리즘 투명성과 책무성에 대한 성명에서 7가지를 제시하였다. 7가지는 인식, 접근과 수정, 책무성, 설명가능성, 데이터 출처 파악, 감가가능성, 유효성 검사와 테스트이다. 2018년 개정 윤리 강령에서 인공지능 윤리 제시. 인공지능 윤리 항목으로 피해의 회피, 투명성과 공정성, 차별에 대한 주의, 개인정보의 존중, 대중인식, 안전을 제시하였고 항목별로 행동 강령을 제시하였다.	강제성은 없는 윤리 강령과 전문가 행동 강령의 일부분으로 제정되었다.
4.Internet Society의 인공지능과 기계학습	인공지능 시스템의 설계와 배포에서 원칙, 윤리적 고려사항과 권고 사항으로 구성되어 있다. 권고사항은 인간의 통제, 안전 최우선, 개인 정보 보호. 데이터 공급의 신중성, 인공지능 시스템의 안전성, 책임 배포이다.	인공지능 시스템의 설계와 배포에서 윤리적 고려사항과 권고 사항을 제기하였고, 배포 과정에서 보안 등을 강조한다.
5.아실로마 인공지능 23원칙	아실로마 인공지능 23가지 원칙은 연구 이슈, 윤리 및 가치, 장기 이슈 3 부분으로 구성되어 있다. 3부분 가운데 윤리와 가치(Ethics and Values)와 직결된 항목들은 13가지이다. 16가지 항목은 안전, 장애 투명성, 사법적 투명성, 책임, 가치 정렬, 개인정보 보호, 자유와 개인 정보, 공동 이익, 공동 번영, 인간의 통제력, 비파괴, 인공지능 무기 경쟁이다.	추상적 수준의 윤리와 가치를 제시하나 세부 행동 준칙 혹은 권고사항을 제시하지 않다.
6.구글 인공지능 원칙	인공지능 관련 비즈니스 의사결정에 영향을 주는 구체적인 기준 7가지를 제시하였다. 7가지 기준은 개인정보보호, 책임, 안전, 편견 조장 금지, 인간의 지시와 통제, 높은 수준의 과학적 수준 유지이다.	인공지능 서비스 제공자의 비즈니스의 사결정에 영향을 주는 선언적 수준의 기준을 제

		시하였다.

부 속 서 A

(본 부속서는 기술보고서 내용의 일부임)

제목

‘해당 사항 없음’

부 록 I

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

제목

‘해당 사항 없음’

부 록 II-1

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

지식재산권 확약서 정보

‘해당 사항 없음’

부 록 II-2

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

시험인증 관련 사항

‘해당 사항 없음’

부 록 II-3

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

본 기술보고서의 연계(family) 표준

‘해당 사항 없음’

부 록 II-4

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

참고 문헌

[1] ACM, ACM Code of Ethics and Professional Conduct, 2018.
[2] European AI Alliance, Ethics Guidelines for Trustworthy AI, April 2019.
[3] IEEE, Ethically Aligned Design, First Edition, April 2019.
[4] Google, 인공지능 원칙 7, 2019.
[5] Internet Society, Artificial Intelligence and Machine Learning: Policy Paper, April 2017.
[6] Future of Life Institute, 아실로마 인공지능 원칙, 2017.

부 록 II-5

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

영문기술보고서 해설서

‘해당 사항 없음’

부 록 II-6

(본 부록은 기술보고서를 보충하기 위한 내용으로 기술보고서의 일부는 아님)

기술보고서의 이력

판수	채택일	기술보고서번호	내용	담당 위원회
제1판	2019.09.09	제정 TTAx.xx-xx.xxxx	-	인공지능 (PG1005)