

군사과학기술정책연구

Military Science & Technology Annual Report

연구논문

• 국방 인공지능 윤리 기준 정립 연구: 최명진, 박헌규



국 방 대 학 교
국가안전보장문제연구소

ISSN 1976-5967

제16권

2023년 12월

군사과학기술정책연구

Military Science & Technology Annual Report

국방대학교 국가안전보장문제연구소

발 간 사

2023년 3월에 국방부에서 「국방혁신 4.0」 기본계획을 발표하면서 국방이 나아갈 방향을 “AI과학기술강군”이라고 하였습니다. 이는 병역자원이 줄어들고 다양한 도전 요소가 있는 미래 국방환경에 대응하기 위해서 유·무인 복합전투 체계를 포함한 미래 전력과 운영에서 과학기술을 통한 혁신을 추진하겠다는 것이고, 그 중심에 AI(Artificial Intelligence, 인공지능)가 있다는 의미로 볼 수 있습니다. 실제로 AI를 국방에 적용하기 위한 다양한 기술개발과 연구가 국방부, 합참, 각 군 그리고 군 관련 기관을 통해서 다양한 분야에서 추진되고 있습니다.

국방의 이러한 방향에 맞추어 국방대학교 국가안전보장문제연구소는 매년 국방부, 합참, 각 군 본부 등의 의견을 수렴하여 군사과학정책 분야에서 필요한 연구주제를 선정하고, 관련 분야의 전문성을 가진 교내·외 연구자들에게 심층적인 연구를 의뢰해 왔습니다. 2023년도에는 국방혁신 4.0이 강하게 추진될 것이었고, 그 중심에 AI가 있었습니다. 국방에서도 다양한 분야에 기술중심으로 AI를 적용하고 있는데, 국방에 AI를 적용할 때 중요한 부분인 AI 윤리에 대한 정책적인 준비가 필요하였습니다. 이를 위해 2023년도 군사과학정책연구주제로 “국방 데이터·인공지능 윤리기준 정립” 연구를 선정하여 연구하고 발간하게 되었습니다.

모쪼록 이러한 연구결과가 국가안보 및 국방정책 관련 부처의 정책개발 및 집행과정에 기여하고, 이 분야를 연구하는 연구자 및 학생들에게 유용한 참고자료로 활용될 수 있기를 기대합니다. 끝으로 제한된 기간에도 불구하고 국가안보를 위한 연구에 열과 성을 다해주신 연구자 여러분께 깊은 감사의 말씀을 전합니다.

2023년 12월 31일

국방대학교 국가안전보장문제연구소장 교수 박 영 준

연구보고 2023

국방 인공지능 윤리 기준 정립 연구

최명진, 박현규

2023. 12.



국방대학교 국가안전보장문제연구소

목 차

요약문	9
1. 서론	25
1.1. 연구목적	25
1.2. 연구범위 및 방법	26
2. 일반적인 윤리와 군사적 측면에서의 윤리	31
2.1. 일반적 개념의 윤리	31
2.2. 군사적 측면에서의 윤리	34
2.3. 소결론	56
3. 인공지능 윤리 이론 및 사례	61
3.1. 인공지능 윤리 관련 이론	61
3.2. 인공지능 부작용 사례(윤리적 측면)	84
3.3. 소결론	95
4. 既 제정된 인공지능 윤리 기준(원칙) 분석	100
4.1. 주요 선진국 및 국내 사례	100
4.2. 정부 부처 및 국내·외 주요 기업 사례	113
4.3. 시사점	123
5. 국방 인공지능 윤리 기준(원칙) 제언	127
5.1. 윤리 기준 가이드 라인 설정	127
5.2. 윤리 기준 핵심 항목 도출	128
5.3. 인공지능 윤리 기준(안) 제언	135
5.4. 인공지능 윤리 기준 평가방안	138
6. 결론	144
붙임	158

그림목차

〈그림 3-1〉 인공지능이란?(인공지능 국가전략, 관계부처합동)	61
〈그림 3-2〉 터키 STM사, KARGU Drone, Dec. 2019.	88
〈그림 3-3〉 국방분야 AI활용에서의 기대 효용 부문 및 발생 가능한 난제	91
〈그림 3-4〉 무기체계 상태진단에 따른 부품 교체 등 최적화 모형	92
〈그림 4-1〉 관계부처 합동(과기정통부), 사람이 중심이 되는 인공지능 윤리 기준	113
〈그림 5-1〉 국방 인공지능 기준 핵심 항목 도출 절차	129
〈그림 5-2〉 10점 척도와 AHP 가중치간의 관계도	134
〈그림 5-3〉 국방인공지능 윤리 기준과 부작용 사례 관계	138

표 목 차

〈표 2-1〉 LAWs 활용에 대한 주요 국가별 견해와 입장	59
〈표 3-1〉 인공지능 윤리 논란 사례	65
〈표 4-1〉 아실로마 AI 원칙	101
〈표 4-2〉 EU의 ALTAI	103
〈표 4-3〉 영국의 데이터 윤리 프레임워크의 중요원칙	104
〈표 4-4〉 영국의 데이터 윤리 프레임워크의 구체적 조치	105
〈표 4-5〉 OECD 인공지능 5원칙	106
〈표 4-6〉 OECD AI 프레임워크 개발을 위한 분류	106
〈표 4-7〉 UNESCO 인공지능 윤리 권고	107
〈표 4-8〉 일본의 거버넌스 가이드 라인	108
〈표 4-9〉 미국의 인공지능 권리장전 청사진	109
〈표 4-10〉 미국 국방성 인공지능 윤리원칙	110
〈표 4-11〉 미국 국방성 인공지능 윤리원칙 권고사항	111
〈표 4-12〉 관계부처 합동(과기정통부), 사람이 중심이 되는 인공지능 윤리 기준 핵심요건	114
〈표 4-13〉 Microsoft 인공지능 원칙	117
〈표 4-14〉 구글의 인공지능 원칙	118
〈표 4-15〉 IBM 인공지능을 위한 일상 윤리	118
〈표 4-16〉 Meta의 책임성 있는 AI 5가지 기준	119
〈표 4-17〉 CISCO의 인공지능 원칙	120
〈표 5-1〉 인공지능 윤리 기준 설문조사 참여 인원	131
〈표 5-2〉 설문조사 결과 우선순위	132
〈표 5-3〉 미 국방성 인공지능 윤리원칙 가중치 분석 결과	135
〈표 5-4〉 국방 인공지능 윤리 기준 점검표 항목 수	140

요 약 문

1. 일반적인 윤리와 군사적 측면에서의 윤리

1.1. 일반적 개념의 윤리

윤리적 원칙들은 상호 연관성이 있으며, 사회적 조화와 공정한 교류를 이끌어 내기 위한 기초 형성 또는 사회적 동의를 위한 거버넌스 구축을 추진하고 있다. 이를 개인, 조직, 사회 수준에서 적용하여 삶의 다양한 영역에서 윤리적인 행동을 실천하는 것이 중요하다. 일반적인 윤리의 기준은 자유주의의 윤리적 개념과 공동체주의의 도덕적 개념을 반영하여 사회 구성원들의 합의에 의해 형성된다고 할 수 있다. 즉, 일반적인 윤리는 사회적으로 공유되고 인정된 규범과 가치를 기반으로 한 행동의 올바름과 옳음에 대한 지침을 나타내는 것이며, 이는 개인, 집단, 조직 또는 사회 전체의 행동 및 결정에 영향을 미치고 상호 존중, 정의, 성실, 책임, 배려, 정직 등의 원칙에 근거한다.

1.2. 군사적 측면에서의 윤리

군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.

1.2.1. 군사용 인공지능의 주요 특성

- 가) 군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.
- 나) 군사용 인공지능 자율무기체계의 개념 및 범위는 군사 로봇에 포함된 자동화 무기(Autonomous Weapons)를 의미하며, 이를 통틀어 치명적인 자동

화 무기체계(LAWs, Lethal Autonomous Weapons Systems) 혹은 자율 살상 무기라 부르기도 한다.

- 다) 군사용 인공지능 자율무기체계의 유형은 자동적 방어체계와 자동적 공격체제로 분류할 수 있으며, 인간이 배제된 신속한 대응을 할 수 있는 체계이다.
- 라) 자율무기체계의 윤리적 쟁점은『위협의 크기, 특성, 양상의 변화 - 기존 위협의 확대, 위협의 특성 변화, 새로운 위협의 출현』과『기계에 의한 의사결정 및 살인의 책임소재가 모호하다는 점』, 『전쟁에 대한 의사결정이 추상적이 될 위협』 측면에서 대두되고 있다.
- 마) 특히 자율살상무기(LAWs)의 윤리적 쟁점은『법적·윤리적 측면』, 『윤리적·도덕적 해이의 문제』, 『생명권 경시의 문제』, 『인간의 존엄과 가치의 문제』, 『인간성의 피폐화와 인간의 수단화 문제』 등으로 대두되고 있다.
- 바) 국제법에 따른 자율무기체계에 대한 윤리적 쟁점은 『국제인도법(군사작전 3대 원칙-구별의 원칙, 비례성의 원칙, 사전주의 원칙)』, 『과잉무기사용금지 원칙』, 『무력충동 원칙 위배 책임』 등으로 대두되고 있다.

1.2.2. 인공지능 자율무기체계와 자율살상무기체계

- 가) 자율무기체계는 물리적 교전, 전자 공격, 사이버 방호 등의 다양한 임무에 활용되며 인간 운용자의 관제를 받는 자율체계까지 포함된다.
- 나) 자율살상무기체계의 구체적인 정의에 대해서는 국제적으로 합의된 바 없으나 자율살상무기는 자율성과 교전권이 부여된 특정 자율무기체계를 나타내는 포괄적인 개념으로 정의한다.

1.2.3. 군사적 측면에서 윤리적 논란

가) 자율살상무기체계 규제를 위한 국제적 논의

- (1) 자율살상무기체계의 개발과 배치, 윤리적 문제를 최초로 제기한 국제기구인 국제 로봇 무기 통제 위원회(ICRAC)로 자율무인전투체계의 연구·개발, 배치, 사용을 금지하는 의제에 대하여 국제적 논의를 제안
- (2) 비정부기구로서 살인 로봇 반대 캠페인(Campaign to Stop Killer Robots)과 국제적십자위원회(ICRC)에서 국제인도법에 위배될 수 있다는 문제점을 제기
- (3) 또한, 국제 인권 감시기구(Human Rights Watch)와 국제 앰네스티(Amnesty International)도 전장에서 활용할 때에 국제인도법을 준수하고 윤리적 행동을 위해 선제적 예방조치가 필요하다고 주장
- (4) 유엔은 2013년 유엔 인권 이사회(UN Human Rights Council)에 제출된 특별보

고서에는 회원국들의 국제적 합의가 있을 때까지 자율살상무기체계의 개발, 생산, 배치를 중지할 것을 권고

- (5) 2014년 유엔의 CCW 회의에서 자율살상무기체계의 개념과 정의, 기술적 특성, 윤리적·안보적 이슈, 국제 규범화 등이 논의되었으나 이에 대한 구체적인 실행 계획이 도출되지 못함

나) 자율살상무기(LAWs) 개발·사용·생산에 대한 국제사회의 동향 및 찬·반론

- (1) 찬성론: 방위비 부담을 줄이고 인명 손실을 최소화하자는 의견, 윤리적 측면에서 꼭 필요한 명령만 내려 로봇 무기가 무고한 살상을 방지할 수 있는 프로그램을 개발하면 피해를 최소화할 수 있는 장점, LAWs에 의한 살인 책임 책임자의 경계는 누가 명령을 내렸는지 기록하기가 더 쉬워 책임자도 명확하게 가려낼 수 있다고 주장
- (2) 반대론: 킬러-로봇이 전투의 문턱을 낮춰 인간의 삶을 크게 훼손할 것이라는 부분, 전 세계적으로 영원한 전쟁 그리고 어떠한 법적 책임이나 민주적 논의도 없는 수많은 비밀 공작 전쟁이 될 것이라고 경고, 근본적인 법적 및 윤리적 문제(이성↔법), 국제법 위배 등으로 반대

1.2.4. 윤리적 논란 해소방안

- 가) 기존 제안된 인공지능 윤리 권고안들은 추상적 원칙들로 이루어져 있어 구체적인 실행 방안이 필요하다.
- 나) 미 국방부가 2020년 2월 발표한 인공지능에 대한 윤리원칙 및 적용하는 가이드 라인이 제시, 한국 군에서도 면밀한 검토를 통해 한국 군사 상황에 맞는 군사 인공지능 윤리 개발 가이드 라인이 필요하다.
- 다) 실질적인 해소 방안으로 윤리적인 설계 및 검수를 거친 군사용 인공지능은 인간의 통제하에 있을 때 오히려 전시에서의 인명 훼손을 이전보다 감소시킬 수 있으므로, 인공지능 무기체계의 기획, 설계, 개발 등의 전체 단계를 통해 윤리적 점검사항들을 정하여 적용할 필요(인공지능 윤리 자율점검 항목들로 구현될 필요가 있음)하다.
- 라) 자율살상무기가 완전한 자율시스템의 수준이 될 때 피아식별이나 전투원-민간인 구분에 있어 인간보다 훌륭한 성능을 보일 수 있는지에 대한 문제이다.
- 마) 가장 높은 단계의 자율성을 가지는 인공지능 수준에서 제기될 문제로 이에 앞서 인간의 제어가 필수적인 반(semi)자율 인공지능에서의 윤리적 설계와 인간의 역할을 구체화할 필요성이 있다.

- 바) 현 한국 군사 관련 현실적 상황 맥락을 검토하여 한국의 이익을 고려하여 수정 및 재구성하는 것이 필요하다.
- 사) 인공지능 편향은 반드시 부정적인 것이 아니며 군사용 무기체계라는 특정 목적을 가지는 자율무기체계에 있어서는 데이터 및 알고리즘 편향이 인공지능시스템 설계 과정에서 자국의 이익을 침해하는 부분과 이익에 도움이 되는 부분을 분리할 필요성이 있으며, 이는 군사적 특성마다 달라지므로 한국의 자국과 적국의 정의 및 그 특성에 따라 구성될 필요가 있다.

2. 인공지능 윤리 이론 및 사례

2.1 인공지능 윤리 관련 이론

2.1.1. 인공지능에서 말하는 윤리

- 가) 인공지능 시스템이 가지고 있어야 하는 윤리적 원칙, 가치, 행동 지침을 의미
- 나) 인공지능은 점점 더 사회 시스템에서 중요한 역할을 수행하며 사회, 경제, 문화, 정치 등 다양한 영역에 영향을 미치게 되면서 그 사용과 개발에 대한 윤리적 고려가 더욱 중요해지고 있는 실정이다.

2.1.2. 인공지능의 윤리원칙

- 가) 공정성과 투명성, 개인정보 보호와 프라이버시, 안전과 보안, 책임과 책임 소재, 사회적 영향과 공익, 자율성과 의사결정 투명성, 사람 중심적 가치, 지속 가능성이다.
- 나) 이러한 윤리적 원칙은 인공지능 개발자, 이용자, 정부, 기업 및 기술 커뮤니티에게 인공지능의 윤리적인 사용과 발전을 촉진하며, 사회적으로 수용 가능한 방향으로 인공지능 기술을 유도해야 한다.

2.1.3. 인공지능 윤리의 필요성

- 가) 인공지능 기술의 급속한 발전으로 제조·의료·교통·환경·교육 등 산업 전 분야에 인공지능이 활용·확산되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 인공지능 윤리 이슈가 새롭게 대두되고 있다.
- 나) 인공지능에 대한 법적 규제와 불충분성으로 규범적 문제에 대응하는 법적

규제가 마련되지 않았다는 점이다.

- 다) 윤리를 통한 자율적 규제는 인공지능 과학기술이 원활하고 창의적으로 발전할 수 있도록 하면서도, 다른 한편으로 인공지능이 유발하는 규범적 문제에 적절하게 대처하기 위해서는 법적 규제보다 직접성이나 강도가 약한 윤리적 규제를 고려할 필요가 있다는 것으로 인공지능과 윤리 문제를 논의할 필요가 있다.

2.1.4. 인공지능 윤리에서 말하는 윤리 개념

- 가) 인공지능 윤리에서 말하는 윤리는 자유주의가 강조하는 도덕 개념으로 한정하여 이해하는 것이 적절하다고 판단된다.
- 나) 현 시점에서 구현되는 있는 인공지능은 약한 인공지능으로 인간에 대한 관계에서 옳음을 준수하는 것으로 인간의 권리나 이익을 침해하지 않는 것이라 할 수 있다.
- 다) 즉, 인공지능에게 요구해야 하는 윤리는 그 의미 내용이 포괄적이면서도 모호한 윤리보다는 외연이 좁으면서도 그 의미 내용이 구체적인 도덕으로 보아야 한다.

2.1.5. 윤리의 구별 개념

- 가) 인공지능 윤리는 공학윤리 가운데 인공지능을 대상으로 하여 전문적으로 특화된 영역이라고 할 수 있으나, 인공지능 윤리는 인공지능을 설계하는 설계자에게 적용되는 윤리뿐만 아니라 인공지능 자체가 준수해야 하는 윤리, 다시 말해 인공지능 자신을 위한 윤리까지 논의 대상에 포함된다.
- 나) 로봇 윤리와 인공지능 윤리를 비교하면 인공지능은 인간의 의식, 즉 의식체계를 인공적으로 구현한 것에 해당되며, 로봇은 인간의 몸 즉, 생명체계를 인공적으로 구현한 것으로 구별되지만, 인공지능과 로봇이 결합되어 등장하는 오늘날의 상황을 고려할 때 양자를 거의 같은 의미로 사용하는 것도 문제가 없을 것으로 판단된다.

2.1.6. 인공지능 윤리의 수범 주체 및 기본방향

- 가) 인공지능 윤리 제정시 수범 주체는 현시점에서 인공지능 개발자 위주의 제정이 우선시 되어야 하며, 향후 인공지능 발전에 따라서 인공지능 자체 및 이용자에 대한 윤리 기준 및 법적인 인정 부분도 고려해야 한다.

나) 결과적으로 인공지능 윤리는 정확하게 말하면 인공지능 개발자가 준수해야 하는 윤리, 인공지능 개발자의 윤리라고 보는 것이 타당하다.

다) 기본방향

- (1) 인공지능 윤리는 자유주의 윤리를 지향해야 하며, 이는 자유주의 윤리의 규범적 의미가 가장 구체적이고 명확하기 때문이다.
- (2) 자유주의 윤리를 가장 우선시 하면서도 보완적으로 공리주의 윤리를 수용할 필요가 있으며, 이는 사람과 사람 사이의 자유나 권리가 충돌하는 상황에서 서로의 이익을 최적화하기 위해서는 이익형량적 사고를 어느 정도 수용해야 함.
- (3) 기본 방향 설정에 따른 윤리원칙: 인간의 존엄성 존중, 권리 침해 금지, 실정법 위반 금지, 이용자의 명령 복종, 투명성, 설명 가능성, 책임

2.2. 인공지능 부작용 사례(윤리적 측면)

2.2.1. 인공지능 투명성, 인공지능 공정성, 인공지능 오작동, 인공지능 보안위협, 기타 부작용 사례 등이 있음.

※ 자율무기체계의 사이버 공격 대응 부재 및 부족, 전쟁 발발 가능성 증대, 자율무기체계의 목적 외 용도로 사용, 생명 경시 및 인간 존엄 훼손 등 이슈 발생 우려

2.2.2. 군사용 인공지능 자율무기체계의 과대적합/과소적합에 따른 문제점

가) 과대 적합·과소 적합은, 학습 반복 수준과 학습 데이터의 분포에 따라 그 결과가 다르게 나타나므로 국방 분야의 특징상, 일부 분야(ex. 전장, 전투 방식 등)에 대한 학습데이터를 확보하기 어렵기 때문에 인공지능 적용 분야를 고려할 필요가 있음.

나) 이러한 현상은 무기체계의 상태 진단 오류, 부품 교체 주기 부적절, 자율무기체계의 미작동 및 오작동, 피아식별 오류 등 발생할 수 있음.

3. 既 제정된 인공지능 윤리 기준(원칙) 분석

3.1. 인공지능 윤리 정책 동향 분석 및 주요 선진국 사례

3.1.1. 인공지능 윤리 정책 동향

- 가) 세계 각국과 글로벌 기업 및 주요 국제기구들은 인공지능의 활용도 증가에 따른 산업의 진흥을 모색하는 한편 윤리적인 인공지능 실현을 위한 정책을 수립해 오고 있다.
- 나) 인공지능 윤리에 대한 필요성을 인식하고 해외 글로벌 기업을 선두로 인공지능 기술에 대한 원칙을 제정하기 시작한 이후 인공지능 학회를 중심으로 각종 이슈를 정리하고 윤리 및 가치를 논의하기 시작했으며 장기적인 관점에서의 원칙들을 발표하기 시작하였다.
- 다) 인공지능 기술과 인공지능 기반 제품 및 서비스가 활성화되고 윤리적 이슈와 개인정보 유출과 같은 문제가 지속 제기됨에 따라 인공지능 선도 국가들은 국가 정책으로 인공지능 활용을 위한 정책을 마련하여 발표하였다.
- 라) 최근 우리나라도 인공지능 윤리 기준 및 원칙을 발표하고 기존에 발표한 인공지능 윤리 기준이 선언적으로만 그치지 않도록 각종 가이드 라인, 개발 기준 등을 마련하고 있다.
- 마) 군사 관련 분야에서는 미국 국방성에서 국제인도법의 원칙을 준수하는 범위 내에서 군사 분야의 특성을 고려하여 국방분야에 적용할 인공지능 윤리원칙을 제시하였다.

3.1.2. 주요 선진국 사례

- 가) 아실로마 AI 원칙은 인공지능 개발과 관련하여 안전, 투명성, 책임, 인간의 가치, 개인정보 보호, 공동 이익 등의 윤리적 기준을 제시하고 있다.
- 나) EU, ALTAI 발표 (20. 7.)
 - (1) 총 143개의 방대한 체크리스트 문항을 포함하여, 산업별·기업별 다른 특성을 가진 기업들이 자신의 상황에 맞게 활용할 수 있도록 작성
 - (2) 자율성을 최대화하고 인공지능이 미칠 수 있는 잠재적인 영향에 대한 인식을 제고
- 다) 영국, 데이터 윤리 프레임워크 발표 (20. 9.)

- (1) 원칙의 적용과 행동 실천을 구체적 점수로 평가하고, 원칙 중 하나라도 3점 이하의 평가를 받을 경우 문제가 있음을 명시하며, 이에 대한 팀내 또는 관련 그룹과의 논의를 통해 해결책을 찾아야 함을 강조
- (2) 실질적으로 실무자들이 확인할 수 있는 수치적 근거를 제시할 수 있도록 하였음
- 라) OECD, 인공지능 신뢰성 확보를 위한 도구 발표 (21. 6.)
 - (1) 개별 도구의 품질이나 완성도는 평가하지 않았으며, 정책 입안자 또는 실무자가 필요한 도구를 신속하게 식별할 수 있도록 참고자료를 제공하는 데 중점을 두고 제작
 - (2) 인공지능 도구를 유형별로 분류·분석한 후, 각 도구들의 정보를 구조화하고, 목적이나 활용 맥락이 서로 다른 도구들을 비교하기 위한 프레임워크 개발
- 마) UNESCO, 인공지능 윤리 권고안 발표 (21. 11.)
 - (1) UNESCO 회원국과 권고안에 언급된 모든 이해관계자들이 해당 권고안에 명시된 윤리적 가치, 원칙을 존중·증진·보호하고 권고 정책의 시행에 필요한 실현 가능한 모든 조치를 취해야 함을 강조
- 바) 일본, 인공지능 원칙 실천을 위한 거버넌스 가이드 라인 발표 (22. 1.)
 - (1) 일본 경제산업성은 본 가이드 라인이 살아있는 문서라는 점에 방점을 두고 있음.
 - (2) 인공지능 기술 발전이 여전히 활발하게 진행되고 있다는 점을 반영하면서, 인공지능의 개발·활용을 사회적으로 수용 가능한 수준에서 관리하기 위함임
- 사) 미국, 인공지능 권리장전 청사진 발표 (22. 10.)
 - (1) 인공지능 권리장전 청사진은 프레임워크로서 기존 미국 내에서 제시되고 있는 인공지능 사용에 관련된 원칙을 전반적으로 정리한 것임.
 - (2) 단순히 원칙만을 제시한 것이 아니라 해당 원칙이 현실에서 실천으로 어떻게 이행될 수 있는지, 미국 연방정부, 국립연구소 등에서 실시하고 있는 각종 연구나 리스크 관리 프레임워크, 관련 표준에 대한 지침과 더불어 민간 사례에 대한 분석을 제공한다는 점에서 의미가 있음.
- 아) 미국 국방성, 인공지능 윤리원칙 발표 (20. 2.)

기본 원칙	주요 내용
책임성	• 적절한 수준의 주의와 판단력을 지속적으로 행사 • (개발, 배치, 사용에 대한 책임도 유지)
공평성	• 인공지능능능에 있어 비의도적인 편향이 최소화되도록 신중하게 조치
추적가능성	• 관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발
신뢰성	• 인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제가능성	• 의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

3.2. 국내·외 사례 분석을 통한 시사점

3.2.1. 해외 글로벌 기업 사례 시사점

가) 인공지능 기술 발전에 따른 부작용 우려

- (1) 인공지능 기술과 서비스가 일상화되고 보편화되는 가운데 잠복해 있던 윤리 훼손 이슈에 대해 관심이 높아졌음.
- (2) 인공지능을 더욱 적극적으로 개발하고 활용범위를 최대한 확대하고자 할수록 인공지능 활용을 통한 이익의 확보와 인공지능에 대한 국민의 보호를 균형적으로 이루어야 함.

나) 인공지능 윤리의 확산

- (1) 인공지능 서비스로 인한 부작용이나 위험을 최소화하기 위해서는 기업의 사회적 책임에 수반되는 윤리의식의 선행이 필요하다는 공감대가 형성됨
- (2) 인공지능의 결함을 해결하지 못하고 인공지능 서비스를 시장에 출시할 경우 엄청난 법적, 재무적 위험에 직면한다는 현실을 직시하게 되었음.
- (3) 이러한 법적, 재무적 손실과 더불어 기업은 기존에 쌓아왔던 명성과 같은 유·무형 손실이 가중될 수 있다는 것을 인지함.

3.2.2. 국내 정부 부처 사례 시사점

가) 인공지능 윤리 기준 제정을 위한 정부의 적절한 개입 필요

- (1) 인공지능 윤리원칙 준수를 위한 기업의 자율적 노력도 필요하지만 관련 정보를 교환하고 자율적 규정을 마련할 수 있도록 정부의 가이드 라인 제정이 필요함

- (2) 인공지능 기술이 적용되는 다양한 범위를 고려하여 위험 수준을 분류하고 고위험 수준의 인공지능 기술에 대해 윤리점검기준을 관리하는 등의 개입이 요구됨
- 나) 인공지능 윤리 준수 우수 기업체에 대한 장려책 제시
 - (1) 인공지능 윤리에 관한 논의가 초기인 점을 고려하여 인공지능 자율점검표 활용이 우수한 기업체를 대상으로 각종 지원을 해 줌으로써 정부 시책을 적극 준수토록 유도할 수 있음
- 다) 인공지능 윤리에 대한 보편적 합의 유도
 - (1) 윤리적 규범의 실천을 위해 인공지능 윤리 관리를 위한 거버넌스 구축과 실천을 촉진할 제도의 마련이 필요함
 - (2) 인공지능 윤리에 관한 이슈에 대응하기 위한 국제적 규범화 노력에 공조할 필요가 있음.

3.2.3. 국내 주요 기업 사례 시사점

- 가) 2018년 이후 인공지능 윤리 관련 문헌 작성 발표
 - (1) 국내 주요 기업들은 인공지능 윤리 헌장, 원칙 또는 준칙, 규정, 점검표를 작성하고 발표하고 있음.
 - (2) 대부분 기업들은 포괄적인 인공지능 윤리원칙을 제시하고 있으나 몇몇 기업들은 자사에서 개발 중인 특정 인공지능 서비스를 대상으로 윤리원칙을 발표함.
- 나) 정부의 인공지능 윤리원칙 부합
 - (1) 기업의 윤리원칙은 대체로 인공지능 서비스를 개발하는 데 있어서 정부의 인공지능 윤리원칙의 핵심 요건에 부합되는 내용으로 발표함.
 - (2) 인공지능 산업과 관련이 있는 국내 주요 기업들은 인공지능 윤리원칙을 발표하고 있지만, 이것의 실천에 관한 공시나 정보를 공개하지는 않음.

4. 국방 인공지능 윤리 기준(원칙) 제언

4.1. 윤리 기준 가이드 라인 설정 및 핵심 항목 도출을 위한 절차

4.1.1. ‘데이터’의 윤리 기준은 ‘인공지능’ 윤리 기준에 포함하여 연구

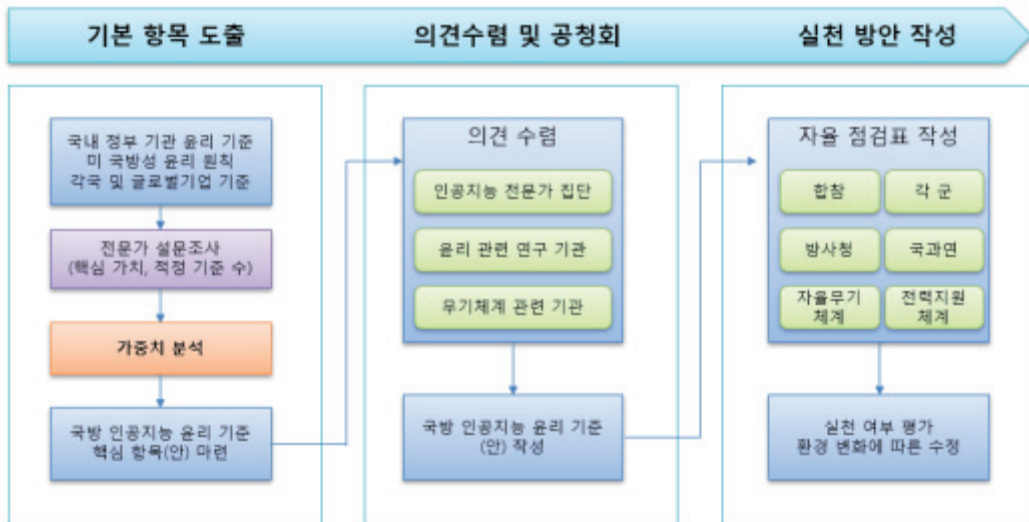
- 가) 데이터의 윤리 기준의 필요성은 있으나 인공지능 윤리 기준의 하위 항목으로 관리
- 나) 별도의 범주로 관리하는 것은 중복소요가 많고 인공지능 윤리 기준과 데이터 윤리 기준을 각각 관리하기가 어려울 것으로 예상됨.

4.1.2. 국내 정부부처 인공지능 윤리원칙과 미 국방성 윤리 기준을 준용

- 가) 각 국과 해외 글로벌 기업 등의 윤리 기준을 살펴보면 유사한 항목들이 많음
- 나) 모든 항목을 고려하는 것보다 많이 준용되는 핵심 항목들 위주로 검토
- 다) 준용할 인공지능 윤리 기준으로 국내 정부부처 인공지능 10가지 원칙과 미 국방성 인공지능 기준 5가지를 참고하여 결정

4.1.3. 설문조사와 가중치 분석 기법을 통해 각 항목별 중요도 평가

- 가) 준용된 인공지능 윤리 기준 핵심 항목에 대해 설문조사로 우선 순위 판단
- 나) 우선순위 판단된 핵심항목에 대해 가중치 분석을 통해 계층적 구조 가시화



4.2. 설문조사 및 가중치 산출 결과

4.2.1. 국방 인공지능 윤리 기준(안) 설정을 위한 항목별 설문조사 결과

- 가) 설문조사 결과 윤리 기준으로서 포함해야 할 항목으로 ‘통제 가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’ 등 4가지는 윤리 기준으로서 우선 순위가 높음을 알 수 있음.
- 나) 추가적으로 ‘추적가능성’, ‘인권보장’도 중요하게 생각하는 요소임을 알 수 있었음.
- 다) 윤리 기준 설정에 따른 키워드(갯 수)로는 6개가 적당하다는 의견이 대부분이었음.
- 라) 따라서 국방분야 인공지능 윤리 기준 우선 순위 설문조사 결과 윤리 기준 6개를 적용시 ‘통제가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’, ‘추적가능성’, ‘인권보장’ 등의 항목이 중요하다는 결과를 알 수 있었음.

구분	1순위	2순위	3순위	4순위	5순위	6순위	기타
항목	통제 가능성	신뢰성	책임성	안전성	추적 가능성	인권보장	공공성, 데이터 관리, 투명성, 침해금지 등
비율(%)	21.4	18.2	17.0	14.5	6.9	6.3	15.7

- 마) 설문조사 제한사항으로 설문조사 항목에 대한 우선순위는 선정할 수 있었으나, 각 항목별 가중치 분석은 실시 할 수 없었으며, 이에 정확한 인공지능 윤리 기준 제정을 위해 핵심 키워드 가중치 분석을 실시함.

4.2.2. 핵심 키워드에 대한 가중치 분석결과

- 가) 가중치 분석은 우선순위 도출을 위해 두 Factor간 쌍대비교 행렬화 연산을 이용한 가중치 산정 기반 의사결정 도구로 평가 항목들을 계층적으로 구성하고 쌍대비교를 통해 문제를 해결할 수 있는 기능을 제공함(AHP 기법 활용)
- 나) 민간 측면에서 도출한 10개 핵심 키워드에 대한 우선순위 및 가중치 산출 결과 아래 그림과 같음.

원칙(핵심 키워드)		순위	10점 척도 (점수비중)		항목	순위	가중치
민 간 측 면	인권보장	3	11.3%		인권보장	5	0.100
	프라이버시 보호	8	8.2%		프라이버시 보호	8	0.064
	침해금지	4	10.8%		침해금지	4	0.102
	다양성 존중	9	5.2%		다양성 존중	10	0.032
	공공성	6	10.5%		공공성	7	0.086
	연대성	9	5.2%		연대성	9	0.036
	데이터 관리	5	10.7%		데이터 관리	6	0.095
	책임성	2	13.2%		책임성	2	0.168
	안전성	1	15.5%		안전성	1	0.211
	투명성	7	9.3%		투명성	3	0.106
총 계			100%		계		1.000

- (1) 1, 2, 4순위는 동일하며, 3순위는 10점 척도에서는 '인권보장', 가중치 분석에서는 '투명성'으로 분석됨.
- (2) 우선순위에서 3순위로 선정된 '인권보장'은 가중치 분석에서는 5순위로 평가됨
- (3) 가중치 분석에서 3순위로 선정된 '투명성'은 10점 척도에서는 7순위로 낮게 평가됨
- (4) 도표의 적색 음영표시는 적용 항목의 경계선을 나타냄
- (5) 민간 측면에서 적용 중인 인공지능 윤리원칙 중 국방 인공지능 윤리원칙에 적용이 가능한 항목은 '책임성', '안전성', '침해금지', '투명성', '인권보장', '데이터 관리'로 분석됨.

다) 미 국방성 인공지능 윤리원칙 항목에 대한 가중치 산출 결과

항목	가중치	내용
책임성	0.235	적절한 수준의 주의와 판단력을 지속적으로 행사 (개발, 배치, 사용에 대한 책임도 유지)
공평	0.056	AI 성능에 있어 비의도적인 편향이 최소화 되도록 신중하게 조치
추적 가능성	0.112	관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발
신뢰성	0.211	인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제 가능성	0.385	의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

- (1) 가중치 분석결과 『통제가능성 > 책임성 > 신뢰성 > 추적가능성 > 공평성』 순으로 중요도가 분석되었으며, 『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단됨

4.3. 국방 인공지능 윤리 기준(안) 제언

4.3.1. 미 국방성 인공지능 윤리원칙 가중치 분석(AHP) 결과

- 가) 『통제 가능성 > 책임성 > 신뢰성 > 추적 가능성 > 공평성』 순으로 중요도가 분석 되었으며, 『공평성』은 신뢰성에 포함될 수 있을 것으로 판단됨.
- 나) 즉, 『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단되어 추후 전문가 의 견수렴 및 『신뢰성』의 정의 설정시 현 미 국방성 인공지능 윤리원칙 『공평성』에 정의 되어 있는 편향성 부분을 반영하는 방안 등 토론(포럼, 세미나 등)을 통해 선정

4.3.2. 민간 측면에서 도출된 항목 중 미 국방성 인공지능 윤리원칙 적용 가능 항목

- 가) 우선순위 및 가중치 분석결과 1·2·4 순위로 동일하게 분석된 핵심 키워드 『안전성, 책임성, 침해금지』는 군사용 인공지능 윤리 기준에 반영 건의
- 나) 민간 측면에서 우선순위 3순위, 가중치 분석에서 5순위로 평가된 『인권보장』 핵심 키워드는 군사적 활용 측면에서 적용 여부는 추가 검토가 필요(포럼, 세미나 등)
- 다) 『안전성』 → 미 국방성 인공지능 윤리원칙의 『신뢰성』에 포괄적으로 적용 가능
- 라) 『침해금지, 인권보장』 → 미 국방성 인공지능 윤리원칙의 『통제 가능성』에 포괄적으로 적용 가능
- 마) 『투명성, 데이터관리』 → 미 국방성 인공지능 윤리원칙의 『추적 가능성』에 포괄적으로 적용 가능 → 국방 인공지능 윤리 기준 적용시 『투명성』으로 적용

4.3.3. 국방 인공지능 윤리 기준 제언

가) 통제 가능성

- 군사용 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생 시 사용자가 그 작동을 제어 또는 정지할 수 있는 기능을 갖추어야 한다.
- 이상 징후 발생시 사용자가 쉽게 인지할 수 있어야 하며 이상 징후에 대한

경고 수준 설정을 할 수 있는 기능을 갖추어야 한다.

- 인간의 의사 결정에 따라서 군사용 인공지능 시스템을 통제할 수 있어야 한다.
- 인간의 통제를 벗어날 수 있는 잠재적 위험 요소 식별을 위해 노력하여야 한다.

나) 책임성

- 군사용 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임 소재를 명확히 할 수 있어야 한다.
- 군사용 인공지능의 사용, 개발, 배포, 결과에 대한 책임을 정의할 수 있어야 한다.
- 인공지능 학습을 위해 수집·저장된 데이터에 대한 관리 책임을 정의할 수 있어야 한다.

다) 신뢰성(안전성, 공정성)

- 외부 공격에 대한 회복성 등 시스템의 안정성을 보장해야 한다. 즉, 군사용 인공지능 시스템의 총수명주기상 성능의 보장, 안전, 보안 등 규정된 범위 내에서 시험되고 보증 되어야 한다.
- 군사용 인공지능의 알고리즘에 대한 공정성과 투명성 보장을 통해 공정한 의사결정(편향성 최소화)을 유도할 수 있어야 한다.

라) 추적 가능성(투명성)

- 군사용 인공지능 알고리즘에 대한 의사결정 과정의 투명성 보장해야 한다. 즉, 군사용 인공지능 시스템은 의사결정 과정을 설명할 수 있도록 보장해야 한다.
- 군사용 인공지능 시스템이 실패, 오작동 일으킬 경우에는 그 이유를 확인할 수 있어야 한다.
- 군사용 인공지능은 설계로부터 전력화되는 모든 개발 과정을 이해관계자와 사용자에게 충분한 정보를 제공할 수 있어야 한다.

마) 데이터 관리

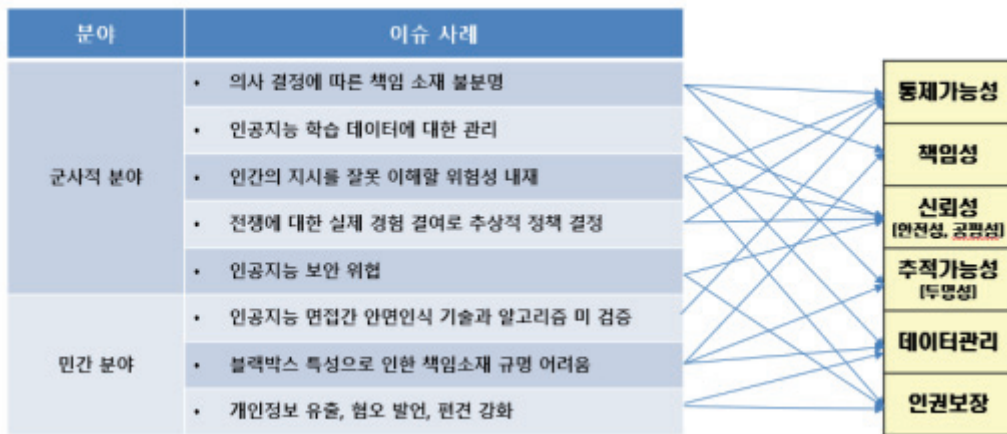
- 군사용 인공지능 개발 및 활용 간 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.

- 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발해야 한다.
- 군사용 인공지능의 편향성(과대·과소 적합) 방지를 위한 데이터 품질 및 위험 관리 프로세스를 마련해야 한다.

바) 인권보장

- 국제인도법을 준수하며 국제법에 명시된 권리를 보장해야 한다.
- 인간을 수단으로 취급되지 않아야 하며, 기본적 권리와 자유를 침해하지 말아야 한다.

※ 군사용 인공지능 무기체계의 부작용 사례와 국방부 인공지능 윤리 기준(안)의 관계



1. 서론

1.1. 연구목적

1.1.1. 배경 및 필요성

- 가) 제4차산업혁명 시대 신기술의 발달, 데이터 기반 행정 효율화 요구 증가, 병역자원 감소 등에 따라 디지털 전환 및 국방지능정보화는 국방 정책·전략의 핵심이슈로 등장하였음.
- 나) 이와 같은 국방환경의 변화에 대응하여 국방분야도 유무인복합체계 구축·운영, 빅데이터 분석을 통한 국방정책의 혁신, 인공지능 과학기술 강군 육성 등 제2차군 수준의 국방혁신4.0을 추진하고 있음.
- 다) 이에 따라 빅데이터 분석 및 인공지능 적용 확대 등을 뒷받침할 수 있는 데이터 및 인공지능 기술의 중요성이 날로 증가하면서 국가적으로는 데이터 기반행정 활성화, 데이터 산업진흥 및 이용촉진, 공공데이터 제공 및 이용활성화 등 데이터 구축과 가치 창출을 위해 노력하고 있으며, 국방분야도 「국방데이터 관리 및 활용 활성화 훈령」 제정 등 데이터 관리 효율화 정책을 적극 추진하는 한편, 인공지능 기술 개발에 투자를 대폭 확대하고 있음.
- 라) 반면, 많은 양의 데이터를 구축하고 이용을 촉진하는 동시에 인공지능 즉, 기계의 지원에 의존한 의사결정이 확대되면서 장애·성별·인종·지역적 차별 문제, 개인정보 유출 문제 등 인권과 사생활을 침해하는 문제들이 발생하고 있어 데이터·인공지능 윤리 기준 정립이 요구되고 있음.

1.1.2. 현실태 및 문제점

- 가) 인공지능은 불가능한 일을 가능하게 하고 새로운 기회를 창출하기도 하지만 정보주체의 사생활 침해, 각종 차별을 강화시키는 등 부작용이 우려되고 있음.
- 나) 반면, 데이터를 수집·활용·제공 시 윤리적 기준에 대한 관심은 초기단계에 머물러 있고, 데이터 윤리를 국방 인공지능·빅데이터 분야에 있어 핵심적인 요소로 고려하고 있지 않기 때문에 이로 인해 발생할 수 있는 문제들에 대한 대비가 부족한 상태임.

다) 인공지능의 경우에는 과학기술정보통신부가「국가 인공지능 윤리 기준(4차산업혁명위원회, '20.12.23.의결)」을 제정하였고, EU·일본 등 주요국가와 마이크로소프트·IBM·카카오 등 국내외 주요 기업에서 인공지능 윤리 기준을 마련하고 있으며 국방분야는 일부 정책연구*가 수행된 상태임.

* 박문언, '군사적 인공지능(AI)의 법적(윤리적) 문제에 대한 검토', 한국국방연구원, 2020

라) 따라서 국방지능정보화를 추진하는 과정에서 데이터를 수집·생성할 때부터 양질의 데이터를 구축하고, 인공지능 및 빅데이터분석 등 데이터 활용 시에도 개인 또는 사회에 부정적 영향을 최소화할 수 있도록 국방 인공지능 윤리 기준을 시급히 마련해야 할 시점임.

1.2. 연구범위 및 방법

1.2.1. 연구범위

연구 방향 설정	
문헌 연구 / 현실태 분석	
- 연구계획 수립 및 연구목표·방향 설정(가정 및 전제조건 설정) - 내부문헌 분석 : 국방부, 육군본부 등 인공지능 무기체계 개발과 관련한 자료 수집 및 분석 - 외부문헌 분석 : 국내외 주요기업의 인공지능 윤리 기준(원칙) 분석	
중요 쟁점 식별 / 분석·평가	
과제 REQ-1, 2	과제 REQ-3, 4 연구수행
(REQ-01) 인공지능 윤리 관련 일반 이론 및 주요 이슈 분석 - 국제적십자사의 자율살상무기체계(LAWS) 논의, 언론보도 내용 포함 (REQ-02) 인공지능으로부터 발생가능한 윤리 관련 부작용 사례조사 - 리비아 내전의 AI 무인기, 우크라이나 AI 안면인식 등 논란 포함	(REQ-03) 주요선진국, 정부부처 및 국내외 주요기업의 인공지능 윤리 기준(원칙) 분석 (REQ-04) 지능정보화 관련 윤리 기준(원칙) 분석 (REQ-05) 국방분야 인공지능 윤리 기준(안)

1.2.2. 연구방법

1.2.2.1. 연구수행 방법 및 절차

가) 중점별 세부과제의 연구결과 도출을 위해 선행연구 및 관련 자료 집중 분석

- (1) 연구중점 사항에서 제시한 5가지 과제에 대하여 문헌조사 연구와 다각적인 의견수렴 및 현장 방문(인터뷰)을 통해 現 실태와 문제점을 진단하고 분석
 - (2) 연구를 진행하면서 필요한 객관적·공식적인 정성적·정량적 자료의 활용
 - 전문가중점인터뷰(FGI) 또는 AHP기법에 의한 연구결과 및 각종 설문조사 결과 등
 - 관련 기관의 공식적이고 공개적인 데이터의 활용 및 진단, 분석
 - 연구기관 및 대외기관의 자료는 인용 또는 참고자료로서 비교 제시
- 요구내용 해결을 위한 연구방향 설정, 사전단계에서 철저한 진단 및 분석

나) 체계적인 연구수행 전략을 수립, 연구성과 달성 극대화 도모

- (1) 효율적이고 수준 높은 연구성과 달성을 위해 치밀한 전략 수립
 - 다양한 정책 연구과제 연구수행 경험과 전문성을 토대로 연구 착수단계에서부터 많은 의견수렴과 교환을 통해 짜임새 있는 연구수행 전략 수립 및 시행
 - (2) 과제 제기부서와의 인식의 차이를 최소화하기 위한 의사소통 활성화
 - 제기부서 담당자와의 정기 / 수시 의견 교환
 - 담당부서의 요구사항을 정확하게 청취/이해, 상호간 의사소통을 원활하게 함으로써 상호 인식의 차이 및 초기부터 연구방향의 편차를 최소화
 - 연구 진행사항과 결과물은 월 단위 제기부서 담당자에게 설명(제출)
 - (3) 무기체계 및 군사학 분야의 전문가로 구성된 연구팀으로 참여자들의 Team Work 제고와 시너지효과로 연구목표 달성
 - 연구 참여자의 Team Work를 높이기 위한 정기/수시 미팅 진행
 - 연구책임자와 자문위원이 자체평가를 실시하여 체계적인 연구 도모
 - 연구 목표 달성을 위해 참여자의 상호 연계성 제고
- 중점과제별 쟁점을 도출, 이를 실현할 수 있는 구체적 방안 마련에 집중

다) 주요 결정 및 영향요소에 대한 진단 및 분석결과 반영

- (1) 새로운 신규소요 판단 및 도출을 위한 일관성과 논리적 전개를 위해 필요
- (2) Top-Down 및 Bottom-Up 접근방법을 병행, 상황 적합적 관점에서 상호순환

- 및 연계시켜 최적의 연구방향 설정하고 주요 과제 및 쟁점 도출
 - 주요 결정요인 및 제반 영향요소를 객관적으로 분석·평가
 - 위협, 전략환경 및 국방환경 변화요소, 전쟁패러다임 변화 등을 반영
- (3) 제기부서 요구내용에 대한 분석평가 및 결과의 논리적 구성과 전개
 - 연구진의 가정 및 전제조건을 기초로 국방부·육본의 관련 계획 및 자료를 분석, 연구 방향을 설정
- (4) 관련 자료 및 객관적 분석을 위해 정성적·정량적 분석을 통해 제시하고자 하는 발전방향과 방안을 최적화할 수 있도록 연구수행
- (5) 자체 설정한 가정과 전제조건을 기초로 국방 인공지능 윤리 기준 정립을 위한 중점과제 및 쟁점을 식별, 분석·평가함으로써 논리적 적절성과 타당성을 개발
 - 새로운 사업의 개발 및 소요 제기를 위한 선택 가능한 방안을 제시
- (6) 신뢰도 높은 윤리적 평가 및 판단요소 도출을 위한 이론적 배경과 논리적 설명이 필요한 부분은 적절한 수준으로 연구내용에 포함

1.2.2.2. 관련기관 협조방안

가) 의견수렴 계획

(1) 관련기관 식별 및 협조방안

군 / 군 관련 기관		기타 관련 기관		협조방안
사업 기관	• 국방부, 국방대학교	연구 기관	• KIDA, ADD, 고등교육기관, 관련 연구소 등	• 전문가중점인터뷰 (FGI) 그룹에 편성 및 자문 • 설문조사
관련 기관	• 방위사업청, 각 군 본부 등			

(2) 협조내용(안)

- 관련기관 방문계획 / 전문가 의견수렴

구 분	내 용
목 적	• 현 인공지능을 활용한 무기체계 적용에 대한 실태 조사(현실태 분석) / 개선방안 수렴
시 기	• 착수보고 후 부터 30일 이내 수시
대 상	① 국방부, 방위사업청, KIDA, 국방대학교, 각 군 관련 업무 부서 ② 필요시 민간 연구기관 등 관련 업체
비 고	• 관련 부서 방문 시 출입조치 및 자료열람 협조 등 제기부서의 지원 필요 • 제기부서의 추천 시 우선적으로 방문 실시

※ 관련기관 및 전문가 의견수렴을 통해 향후 연구내용에 대한 객관성과 전문성을 확보하여 연구의 신뢰도를 향상시킬 예정임.

→ 국방 인공지능 윤리 기준 정립을 위한 현실태 분석 및 개선방안에 대한 의견 수렴 후 표준안으로 활용할 수 있도록 연구결과 제시

1.2.3. 연구과제 관리

가) Step ❶ 요구사항을 충족할 수 있는 연구계획서 작성으로 연구수행 위험(Risk) 최소화

- 제기부서의 요구사항 분석을 통해 초기부터 과제 수행 가능한 연구 계획서 제출



나) Step ❷ 과제수행 검토회의 개최, 선행연구 및 현실태 진단분석

- 제기부서 '착수보고' 이전에 연구계획서 재검토(수행계획 전반을 평가하여 '착수보고 준비')
- 문헌조사와 현장 방문 및 인터뷰를 통한 다양한 자료 수집
- 사례분석, 중점과제별 세부과제 도출 및 주요 쟁점 설정

다) Step ❸ 과제 중간검토회의 개최(1~2차, 현장 인터뷰 병행)

- 제기부서 '중간보고' 2주전에 '과제 중간 검토회의'를 통하여 연구 전반에 대한 종합 검토, 수행 간 문제점 식별

- 중점-핵심 세부과제의 집중 분석·평가
 - 연구방향에 따른 과제별 주요 발전방향 및 다양한 방안 구상, 선정
 - 1차 중간보고를 통해 연구방향 조정 및 결정
 - 1차 중간보고를 통해 결정된 연구방향을 확정하고 심층 깊은 연구실시
 - 2차 중간보고 실시 / 지속 보완, 외부 전문가 자문 및 토론회
- 라) Step ④ 과제 최종 검토회의 개최
- 제기부서‘최종보고’1개월 전에‘과제 최종 검토회의’를 통하여 제기부서 요구사항 충족 여부 및 연구결과 종합 검토/평가
- 마) Step ⑤ 최종 연구보고서 Review
- 연구보고서 초안(Draft)을 검토위원이 제기부서 입장에서 종합 검토하여 연구보고서 완성도를 높이도록 함
 - 과제별 다양한 방안의 비교분석 및 최적의 방안 도출 / 최종보고 실시 / 최종보완 제출
- 제기부서의 요구내용을 중심으로 분석 및 연구를 수행하되, 종합적으로 중점 과제 및 세부과제 단위별로‘상호 연계성과 실현 가능성’에 주안을 두고 단계적·반복적 분석 및 조정을 거쳐 방안 및 정책 방향 제언

2. 일반적인 윤리와 군사적 측면에서 윤리

2.1. 일반적 개념의 윤리

2.1.1. 개요

일반 인공지능 윤리는 인공지능이 사회 전반에 적용될 때 지켜야 할 윤리적 원칙이며, 군사용 인공지능 윤리는 군사 분야에 특화된 인공지능 기술과 관련하여 지켜야 할 윤리적 원칙이라고 할 수 있다. 일반 인공지능 윤리에서는 인공지능이 인간의 권리와 자유를 침해하지 않도록 개인정보 보호, 투명성, 책임성 등을 강조하는 반면, 군사용 인공지능 윤리는 군사 목적을 위한 인공지능 개발과 사용에 있어서의 투명성(추적관리성), 책임성, 신뢰성(안전성) 등을 강조하고 있다. 또한, 군사적 목적을 위해 인공지능을 사용하는 과정에서 국제인도법 등 국제법과 관련하여 인간의 생명과 존중, 안전을 고려해야 한다는 점이 이슈화 되고 있는 상태이다. 그러나 군사용 인공지능은 적에게 탐지되지 않도록 스텔스 기능을 강화하거나, 인간이 수행하기 어려운 위험한 임무를 대신 수행하는 등의 목적으로 개발될 수 있기 때문에 일반 인공지능에서는 허용되지 않는 일부 윤리적 제약이 완화될 수 있다. 즉, 인간의 존엄성과 인권을 침해하지 않도록 적절한 규제와 감시가 필요한 상태이다.

따라서, 군사용 인공지능 윤리는 일반 인공지능 윤리를 바탕으로 군사 분야의 특수성을 고려하여 추가적인 원칙과 규제가 필요하고, 이를 통해 인공지능 기술이 군사 분야에서도 인간의 존엄성과 인권을 존중하며, 안전하게 사용될 수 있도록 해야 한다.

2.1.2. 일반적인 윤리란?

일반적인 윤리란 사회 구성원들이 지켜야 할 보편적인 도덕적 가치와 규범을 말하며, 일반적인 윤리는 시대와 문화에 따라 변화할 수 있으며, 사회 구성원들의 합의에 의해 형성된다. 이를 통해 사회는 질서를 유지하고, 공정한 사회를 구현할 수 있는 것이다. 즉, 일반적인 윤리는 사회적으로 공유되고 인정된 규범과 가치를 기반으로 한 행동의 올바름과 옳음에 대한 지침을 나타내는 것이며, 이는 개인, 집단, 조직 또는 사회 전체의 행동 및 결정에 영향을 미치고 상호 존중, 정의, 성실, 책임, 배려, 정직 등의 원칙에 근거한다.

2.1.2.1. 일반적인 윤리원칙과 가치

- 가) 진실성과 정직: 다른 사람에게 솔직하고 정직하게 행동하며, 거짓말을 하지 않고 타인을 속이지 않는다.
- 나) 책임과 의무: 자신의 행동에 대한 책임을 집니다. 타인에 대한 의무를 이행하고, 일반적인 사회적 규범과 법에 따라 적절하게 행동한다.
- 다) 공정과 정의: 모든 사람에 대해 공평하고 공정하게 행동하며, 차별이나 편견
- 라) 존중과 배려: 다른 사람의 인권과 존엄성을 존중하고 배려하며, 각 개인의 견해와 가치를 존중한다.
- 마) 투명성: 투명하고 공개적으로 행동하며, 정보와 결정에 대해 투명성을 유지하여 타인이 신뢰할 수 있도록 한다.
- 바) 자기결정권과 자율성: 개인의 자기결정권을 존중하고, 다양한 의견과 가치를 인정하여 개인의 자율성을 존중한다.
- 사) 안전과 건강: 자신과 다른 이의 안전과 건강을 보호하고 유지하기 위해 노력하며, 안전한 환경을 조성한다.
- 아) 환경보호: 환경을 보호하고 지속 가능한 방식으로 자원을 사용하며, 생태계에 대한 책임을 가진다.
 - ✓ 윤리적 원칙들은 상호 연관성이 있으며, 사회적 조화와 공정한 교류를 이끌어내기 위한 기초 형성 또는 사회적 동의를 위한 거버넌스 구축을 추진하고 있다.
 - ✓ 이를 개인, 조직, 사회 수준에서 적용하여 삶의 다양한 영역에서 윤리적인 행동을 실천하는 것이 중요하다.

2.1.2.2. 윤리와 도덕의 관계

우리가 일상적으로 흔히 사용하는 도덕 개념과 어떤 관계를 맺는지, 양자는 같은 개념인지, 아니면 서로 구별되는 별개의 개념인지에 따라 그 의미 내용이 달라진다. 이는 철학적으로 자유주의를 취하는가, 아니면 공동체주의를 취하는가에 따라 달라진다.

예를 들면 개인과 공동체의 관계에서 개인의 자유를 공동체의 미덕보다 우선하는 자유주의는 철학적으로 볼 때 무연고적 자아(une-ncumbered self) 개념, 좋음(the good)에 대한 옳음(the right)의 우선성이다.¹⁾

1) 양천수(2014), 자유주의적 공동체주의의 가능성: 마이클 샌델의 정치철학을 중심으로 하여, 『법철학연구』, 제17권 제2호, pp. 205-242.

가) 자유주의의 윤리적 개념

- (1) 자유주의는 공동체 또는 사회적 관계로부터 자유로운 선험적 주체, 즉 무연고적 자아 개념을 인정하는 것으로 자아는 공동체가 성립하기 이전에도, 사회적 관계가 형성되기 이전에도 존재한다는 것이다.
- (2) 자유주의는 좋음에 대해 옳음이 우선한다고 본다. 이러한 주장은 좋음과 옳음은 개념적·내용적으로 구별된다는 것이다.
- (3) 좋음은 공동체의 미덕과 관련되는 것이며, 반면에 옳음은 무연고적 자아인 개인과 개인의 관계에서 문제되는 것이기 때문이다.
- (4) 이러한 맥락에서 자유주의는 윤리와 도덕을 개념적·내용적으로 구분한다.

나) 공동체주의의 도덕적 개념

- (1) 윤리가 좋음 및 공동체의 미덕과 관련되는 것이라면, 도덕은 옳음 및 자아 사이의 관계 설정 문제와 관련된다.²⁾
- (2) 그러면서 도덕이 윤리보다 우선한다고 말한다. 이에 반해 개인을 공동체를 넘어서는 존재로 보기보다는 공동체 안에 있는 존재로 파악하는 공동체주의는 철학적인 면에서 자유주의와 대립되는 주장을 한다.
- (3) 자아에 관해서는 연고적 자아(encumbered self)를, 좋음과 옳음의 관계에 대해서는 좋음이 옳음에 우선한다고 말한다.
- (4) 공동체주의는 자유주의와는 달리 자아는 공동체 또는 사회적 관계와 무관하게 존재할 수 없다고 말한다. 모든 자아는 연고로 대변되는 사회적 관계에서 자유로울 수 없다는 것이다.
- (5) 즉, 공동체주의는 좋음과 옳음을 구별하면서 옳음에 우선권을 부여하는 자유주의에 반대되며, 일단 공동체주의는 좋음과 옳음을 내용적으로 구별할 수 없다고 말한다. 옳음의 배후에는 좋음이 놓여 있기 때문이다.
- (6) 결과적으로 옳음 그 자체도 좋음의 한 유형에 불과하다는 것으로 곧 좋음이 옳음에 우선한다는 것을 의미한다.
- (7) 이와 같이 공동체주의는 도덕과 윤리를 내용적으로 구분하는 것에 반대한다. 개인적 도덕의 배후에는 공동체적 윤리가 있다는 것이다. 이러한 공동체주의의 견지에서 보면, 윤리는 (이보다 그 외연이 좁은) 도덕 개념을 포괄하는 개념으로서 옳음뿐만 아니라 좋음까지 다루는 규범들의 총체라 할 수 있다.

2) 독일의 사회철학자 하버마스(Jürgen Habermas)는 실천이성의 도덕적 사용과 윤리적 사용을 구분한다. 위르겐 하버마스, 이진우(역)(1997), *담론윤리의 해명*, 문예출판사.

- ✓ 따라서 일반적인 윤리의 기준은 자유주의의 윤리적 개념과 공동체주의의 도덕적 개념을 반영하여 사회 구성원들의 합의에 의해 형성된다고 할 수 있다.
- ✓ 즉, 일반적인 윤리는 사회적으로 공유되고 인정된 규범과 가치를 기반으로 한 행동의 올바름과 옳음에 대한 지침을 나타내는 것이며, 이는 개인, 집단, 조직 또는 사회 전체의 행동 및 결정에 영향을 미치고 상호 존중, 정의, 성실, 책임, 배려, 정직 등의 원칙에 근거한다.

2.2. 군사적 측면에서의 윤리

2.2.1. 군사용 인공지능의 특성 및 윤리적 측면의 논의

군사용 인공지능은 군사 분야에서 활용되는 인공지능 기술 및 시스템으로 다양한 군사 작전, 안보, 정보 수집, 전략 수립, 장비 관리, 교육훈련 및 의사결정 지원 등에 적용되어 군사력 및 효율성 향상을 목표로 하고 있다.

2.2.1.1. 군사용 인공지능의 주요 특성

군사용 인공지능은 군사 분야에서 중요한 역할을 하지만, 윤리적인 문제가 발생할 수 있다. 예를 들어, 무인 무기 시스템이 오작동하여 민간인을 공격하거나 군사 정보 분석 결과가 잘못된 판단을 내리는 등의 문제가 발생할 수 있다. 따라서 군사용 인공지능 개발과 사용에 있어서는 윤리적 측면에서 검토할 부분이 있으며, 이에 다양한 국제기구와 사회단체, 관련 글로벌 기업에서는 군사용 인공지능 활용에 따른 윤리적인 가이드 라인을 제정하고 있는 추세이다.

가) 군사용 인공지능을 활용

- (1) 자율적 무기 시스템: 인공지능은 무인 기체, 로봇, 드론과 같은 자율적 무기 시스템을 통해 탐지, 타격, 정찰 등의 작전을 수행합니다.
- (2) 지능형 정보 수집 및 분석: 군사용 인공지능은 풍부한 데이터를 분석하여 정보 수집, 분석, 정리를 효율적으로 수행하고 전략적인 통찰을 제공합니다.
- (3) 사물인터넷(IoT) 기반 통신 및 네트워킹: 인공지능을 활용한 IoT 시스템은 군사 장비와 플랫폼 간에 효율적인 통신을 가능케 하며, 실시간 정보 교환을 강화합니다.
- (4) 전투 시뮬레이션 및 훈련: 군사용 인공지능은 가상 혹은 현실의 전투 시뮬레이션

션을 통해 군인의 훈련과 전투 시뮬레이션을 지원하며 전략적인 훈련 효과를 높입니다.

- (5) 긴급 상황 대응 및 의사결정 지원: 인공지능은 긴급 상황에서 빠르고 효율적인 대응을 위해 의사결정 프로세스를 지원하고 필요한 정보를 신속하게 제공합니다.
- (6) 예측 분석과 전략 수립: 데이터 기반의 예측 분석을 통해 전략 수립을 지원하고 전술적으로 중요한 지점을 파악하여 군사 작전을 계획합니다.

✓ 군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.

2.2.1.2. 인공지능 자율무기체계의 개념 및 범위

가) 인공지능 자율무기체계(AWS, Autonomous Weapon Systems)란?

- (1) 인공지능 기술을 활용해 스스로 목표물을 추적·공격할 수 있는 공격용 전투 로봇을 포함한 인공지능의 무기체계를 의미하며, 일반적인 무기체계와 달리 인공지능기술을 활용해 목표물을 추적·공격할 수 있는 것이 특징이다.
- (2) 넓은 의미에서 군사 로봇(Military Robot)에 포함된 자동화 무기(Autonomous Weapons)를 의미하며, 이를 통틀어 치명적인 자동화 무기체계(LAWs, Lethal Autonomous Weapons Systems) 혹은 자율 살상 무기체계라 부르기도 한다.
- (3) 사람의 개입 없이 스스로 판단해 공격할 수 있다면 킬러-로봇으로 분류하며, 자율무기체계에서 자율성은 무엇보다 인공지능과 밀접한 관련이 있다.
- (4) 자율무기체계의 자율적 작동이 인공지능에 기반한 알고리즘을 통해서 이루어질 수 있기 때문이며, 자율무기체계에서 핵심은 자율성의 정도이다.
- (5) 자율성에 관해 합의된 개념 정의는 없지만 기본적으로 이것을 실체적 개념이 아니라 관계적 개념으로 이해할 필요가 있음. 즉 인간-기계 간 지휘-통제 관계에서 인간이 관여하는 정도로 자율성을 이해할 수 있다는 것이다.³⁾

나) 윤리적 측면에서 킬러-로봇의 개발을 제한

- (1) 핵무기보다 비용이 저렴하고, 대량 생산이 가능해 분쟁 지역 등에서 악용될 수 있기 때문에 이런 이유로 2013년에는 노벨 평화상 수상자인 조디 윌리엄스와 인권단체 등이 킬러-로봇 반대 운동을 조직하였다.

3) 조현석, “인공지능, 자율무기체계와 미래 전쟁의 변환”, 『21세기정치학회보』 28(1), 2018. 3, pp. 115~139, p.118.

- (2) 2015년에는 스티븐 호킹(Stephen Hawking)과 애플 공동 설립자 스티브 워즈니악(Steve Wozniak)을 비롯해 1,000여 명의 인공지능·로봇 공학 연구자들이 인공지능 관련 국제회의(International Joint Conference on Artificial Intelligence)에서 인공지능 전쟁 금지를 권고하는 서한에 서명하기도 했으며, 이 서한에는 인공지능 로봇을 군사 무기로 활용하는 킬러-로봇의 위험성을 담고 있다.

다) 기타 활용 범위

- (1) 킬러-로봇과는 별도로 여러 군사 로봇이 개발되어 현장에서 활용되고 있으며, 로봇 제조사 보스턴 다이나믹스(Boston Dynamics)에서 개발한 사족보행 로봇 알파 독(Alpha Dog)의 경우 보행성능과 수송용량 등 기능을 개선해 LS3(Legged Squad Support Systems)라는 이름의 군사 로봇으로 쓰일 예정이며, LS3는 거친 지형에서 장비와 연료 등을 운반하고 GPS를 사용해 지정된 위치로 이동할 수 있다.

2.2.1.3. 인공지능 자율무기체계의 유형

가) 자동적 방어체계(Automatic defensive systems)

- (1) 자동으로 작동되는 군사 무기로 가장 오래된 치명적인 무기는 1,600년대 이후로 사용된 육군 지뢰와 1,700년대 이후부터 사용된 해군 지뢰이다.
- (2) 그러나 현재의 자율적 방어 무기체계의 예로는 1970년대 이래 미국 해군이 선박을 방어하기 위해 사용하는 Phalanx CIWS 레이더 유도 총과 자동화된 "하드 킬" 액티브 보호 시스템이 있다.
- (3) 이러한 방어적 자율무기 시스템은 조작자가 설정한 기준에 따라서 접근해 오는 미사일, 로켓, 포병 화기, 항공기와 군함으로부터의 공격을 자동적으로 식별하고 자율적으로 반격할 수 있도록 설계되어 있다.
- (4) 이스라엘이 실전에 배치하여 운영하는 아이언 돔(Iron Dome) 미사일 방어시스템은 자율적 표적화 능력을 갖추고 있으며, 러시아 아레나(Russian Arena), 이스라엘 트로피(Israeli Trophy), 독일 AMAP-ADS와 같은 시스템이 유사한 것으로 탱크에 배치되어 있다.
- (5) 군용차량에 설치된 자동 포탑을 원격무기 스테이션이라고 하며, 이러한 자율적 무기 시스템에서 인간이 배제되는 주된 이유는 신속한 대응의 필요성 때문이다.⁴⁾

4) "Getting to grips with military robotics.", The Economist, 25 January 2018.

나) 자동적 공격체계(Autonomous offensive systems)

- (1) 오늘날 미군이 운용하는 BAE Systems Taranis, 일명 랩토(Raptor)와 같은 무인 자율 제트추진 전투기는 자발적으로 적을 찾고, 식별할 수 있는 미래형 공습 시스템이지만 현재도 공격 명령만 입력되면 표적을 공격하고 교전하며 적기의 공습도 방어할 수 있는 능력을 갖춘 것으로 알려져 있다.
- (2) 항공모함에서 자동 이착륙하는 Northrop Grumman X-47B 무인 항공기는 자율적 출격과 감시 및 공격(UCLASS, Unmanned Carrier-Launched Airborne Surveillance and Strike)시스템으로 개발되었다.
- (3) 2018년 미국의 핵 태세 검토보고서(U.S. Nuclear Posture Review)에 따르면 미국은 러시아가『Status 6』로 명명된 신형 대륙간 핵 추진 해저 자율어뢰를 개발 중인 것으로 파악하고 있다.
- (4) 그 밖에도 러시아는 인공지능 미사일, 무인항공기, 무인군사차량, 군용로봇 및 의료용 로봇을 적극적으로 개발하고 있다.
- (5) 2017년 이스라엘 Ayoub Kara 국방장관은 이스라엘이 헤즈볼라와 하마스 지도자들을 암살하기 위해 파리와 같이 매우 작은 군사용 로봇을 개발하고 있으며 이 로봇은 최소 3년 이내에 실전 배치될 것으로 예상된다.⁵⁾

2.2.1.4. 자율살상무기체계(LAWs)의 정의

가) 자율무기체계(LAWs, Lethal Autonomous Weapon systems)란?

- (1) 자율살상무기체계는 교전권과 살상권이 부여된 자율무기체계로서 인간 운용자나 통제자의 개입 없이 스스로 표적을 탐색하고 선택하여 공격하거나 교전할 수 있는 무기체계를 말한다.
- (2) 자율살상무기체계와 자율무기체계는 자주 혼용되어 사용되는데 자율무기체계(AWS, Autonomous Weapon Systems)는 자율살상무기체계보다 상위의 개념으로서 외부 환경을 자체적으로 평가하고 선택 가능한 행동 방책의 범주 내에서 다음 방책을 자율적으로 선정하며, 환경의 역동적인 변화에도 신속하게 적응할 수 있는 체계로 정의할 수 있다.
 - ✓ 자율무기체계는 물리적 교전, 전자 공격, 사이버 방호 등의 다양한 임무에 활용되며 인간 운용자의 관제를 받는 자율체계까지 포함
 - ✓ 자율살상무기체계의 구체적인 정의에 대해서는 국제적으로 합의된 바 없으나 자율살상무기체계는 자율성과 교전권이 부여된 특정 자율무기체계를 나타내는 포괄적인 개념으로 정의

5) "US says Russia 'developing' undersea nuclear-armed torpedo". CNN. 2018.

나) 자율무기체계(AWS, Autonomous Weapon Systems) 활용 범위

- (1) 자율살상무기체계는 일반적으로 무인전투체계로서 자동로봇과 형상이나 운용방식이 유사하여 종종 자율 살상 로봇(Lethal Autonomous Robots), 로봇형 무기체계(Robotic Weapons), 킬러-로봇(Killer Robots) 등으로도 알려져 있다.
- (2) 정의에 따르면 현재까지 구현된 자율무기체계 중에서 자율살상무기체계로 분류할 수 있는 무기체계는 현존하지 않는다고 볼 수 있으나 이에 근접한 무기체계로서 선회체공탄(Loitering Attack Munition)⁶⁾을 들 수 있는데 이 체계는 특정 지역으로 발진하여 사전에 프로그램된 기준에 따라 레이더기지, 전투함정, 전차와 같은 표적을 탐색하고, 식별된 표적에 대해서는 공격으로 전환하여 파괴하는 무기체계이다.
- (3) 대표적인 예로서 이스라엘에서 개발하여 운용 중인 무인공격기 하피(Harpy)⁷⁾를 들 수 있으며, 하피는 표적의 선정과 공격에 있어 일부 자율성이 부여되어 있으나 의사결정 과정에서 인간 운용자의 감독과 통제를 받기 때문에 완전한 자율살상무기체계로 보기는 어렵다.
- (4) 이와 유사한 반자율(Semi-Autonomous) 수준의 또 다른 무기체계로서 영국의 타라니스(Taranis) 드론, 미 해군의 자율함정 시헌터(Sea Hunter), 보잉의 무인잠수정 에코 보이저(Echo Voyager), 러시아의 무인전차 우란-9(Uran-9) 등이 있다.⁸⁾

2.2.1.5. 자율무기체계의 윤리적 쟁점

가) 위협의 크기, 특성, 양상의 변화

- (1) 기존 위협의 확대: 인공지능시스템을 활용한 효율적인 공격이 더 쉬워지고 공격 속도도 증가함. 또한 특정인이나 단체를 대상으로 하여 공격대상자에 대한 정보를 수집 분석하여 피싱공격을 수행하는 스피어피싱(spear phishing)이 대두되었다.
- (2) 위협의 특성 변화: 인공지능을 이용한 공격의 빈도와 효과 및 위협의 크기가 동시에 모두 증대하였다.
- (3) 새로운 위협의 출현: 인공지능에는 시공간적 제약이 없으므로 인간이 할 수 있

6) 조현식, “인공지능, 자율무기체계와 미래 전쟁의 변환,” 『21세기정치학회보』 제28집 제1호, (2018), p. 124.

7) 무인공격기 하피는 Israel Aerospace Industries에서 생산된 무기체계로서 레이더 시스템을 공격하거나 방공무기체계를 제압하기 위해 설계되었다. 하피는 전개된 지역을 일정한 패턴에 따라 선회 체공하다가 식별된 표적에 대해서는 장착한 미사일을 ‘fire-and-forget’ 방식으로 발사하여 파괴한다.

8) 조현식, “인공지능, 자율무기체계와 미래 전쟁의 변환,” 『21세기정치학회보』 제28집 제1호, (2018), pp. 123~124

는 것뿐만 아니라 인간이 할 수 없는 공격 수행까지 가능하고 드론처럼 테러와 암살에서 공격대상자와 대면하지 않아도 공격을 할 수 있는 새로운 위협이 등장하였다.

나) 기계에 의한 의사결정 및 살인의 책임 소재가 모호하다는 점

- (1) 특히 과거의 인공지능 모형이 아닌 딥러닝 방식으로 설계되는 인공지능은 데이터가 학습되는 방식이 블랙박스의 특성을 가지고 있어서 학습되는 데이터들이 어떻게 패턴 파악되는지의 과정을 알 수 없다.
- (2) 따라서, 인공지능이 내린 의사결정이 합리적으로 보이더라도 그에 대한 정당화 작업이 어려우며, 합리적 의사결정의 동기가 잘못된 것일 가능성을 배제할 수 없다.
- (3) 동일 맥락에서, 인간 지시가 병행되는 반자율 인공지능을 사용하더라도 인간지시를 잘못 이해할 위험성이 있다.

다) 전쟁에 대한 의사결정이 추상적이 될 위험

- (1) LAWs가 완전 자율무기체계(FAWs: Fully Automated Weapons system) 체계가 될 경우 전장에서 사람을 배제시키게 됨으로써 전쟁에 대한 실제 경험을 결여하게 되어 추상적 정책결정을 하게 된다.
- (2) 인공지능 무기체계가 인간의 제어하에 윤리적으로 사용된다고 할지라도 실제 전쟁터에서는 살상 대상을 인간으로서 존중하는 현실은 존재하지 않는다.
- (3) 따라서 인간의 수단화를 쟁점으로 고려하기보다는 인공지능의 정확한 피아식별 성능의 향상을 모색하는 것이 현실적이다.

2.2.1.6. 자율살상무기(LAWs)의 윤리적 쟁점

인간의 개입 없이 자율적으로 살해 결정을 내리는 인공지능 LAWs이 국제협약에 의해 보호되는 생명권이나 인도법칙을 위반하는지 여부가 중요하지만⁹⁾ 인공지능 LAWs에 대한 국제 보편적 개념정의가 없는 상태에서 각국이 국가안보 관점에서 국제 관련 규정을 해석하는 현재 환경에서는 국제무기규범의 적용은 한계가 있다.

가) 법적·윤리적 쟁점의 등장

- (1) 이론적인 공방에도 불구하고 실전의 인공지능 무기체계 개발 주체들은 논란을

9) 인공지능(AI) 치명적자율무기체계(LAWs)의 법적·윤리적 쟁점에 대한 기초연구, 중앙법학회, 한의원, 2018. 3. pp. 325~365.

크게 신경 쓰지 않는 듯하지만 비난 여론은 여전한 상태이며, 이는 프로그램 알고리즘을 활용한 지능형 기계는 언제든지 자체 판단하는 인명 살상용 무기체계로 전환될 수 있다는 우려 때문이다.

- (2) 전문가들은 2025년 정도에는 무인 자율로봇이 전쟁의 중심에 등장할 것이라고 예고하고 있으며, 이는 『킬러-로봇』을 반대한 일론 머스크가 지적한 바와 같이 “인공지능 무기체계가 핵무기보다 싸고 대량 생산이 가능”하기 때문이다.¹⁰⁾
- (3) 전문가들은 몇 년 안에 인공지능 자율무기체계의 배치가 가능한 시점에 도달했다고 경고하고 『테러리스트나 독재자가 킬러-로봇을 무고한 사람들에게 사용할 수 있고, 킬러-로봇이 해킹된다면 무자비하게 쓰일 수 있다』고 우려하면서 『일단 이 판도라의 상자가 한 번 열리면 닫기 어려울 것』이라고 말했다.
- (4) 킬러-로봇에 의한 윤리적 법적인 의문
 - ① 만일 로봇이 사람을 죽인다면 그 책임은 누가 져야 할까? 도덕적 책임에 그칠 것인가? 법적인 책임을 추궁할 수 있을까? 영화에 자주 등장한 것처럼 인공지능 회로에 문제가 생기면 통제 불능이 된다.
 - ② 통제 불능이 된 자율무기체계는 피아식별을 제대로 하지 않을 뿐만 아니라 민간인과 군인도 구별하지 못하고 무차별적으로 공격에 나설 수 있으며, 적대국이 아직 선전포고를 발령하지도 않았는데 사전공격을 하는 것으로 오인하고 선제공격을 단행하고 급기야 전면전쟁을 야기할 위험성도 부인할 수 없다.
 - ③ 로봇의 오류는 누가 책임지나? 과연 로봇에게 인간을 죽일 수 있는 권한을 부여할 수 있을 것인가? 만일 킬러로봇이 오·작동되어 적군 대신 아군을 대량 살상한다면 그 책임을 누구에게 물을 것인가?

나) 윤리적·도덕적 해이의 문제

- (1) 인공지능전문가인 스튜어트 러셀(Stuart Russell)은 LAWs가 반윤리적이고 비인간적이라면서 이 시스템의 가장 큰 문제점은 전투원과 비전투원을 구별하는 것이 어렵다는 점이다.¹¹⁾
- (2) 또 다른 위험으로 원격 제어 무인 항공기 파업과 마찬가지로 LAWs는 일부 당사자에게 전쟁을 보다 쉽게 생각하고 자행하여 더 많은 살상을 초래할 수 있다는 점이다.
- (3) 완전자율무기체계(Fully Automated Weapons: FAWs) 체계는 전장에서 사람을 배제시키게 됨으로써 군인과 일반 시민, 정치인과 정책결정권자가 실제 전

10) 중앙일보, 북핵보다 치명적인 인공지능 무기, 2018. 02. 17

11) Russell, Stuart(27 May 2015). "Take a stand on AI weapons". International weekly journal of science. 521.

장 터의 참상을 직접 경험할 기회를 배제하고 추상적이고 일반적인 기대와 희망에 기초하여 정책결정을 하게 함으로써 전쟁의 문턱을 낮추고 전쟁으로 가는 자동화된 도로를 열어 줄 수도 있을 것이라는 점이다.

- (4) 또한, FAWs가 일단 실전에 배치되면 전쟁에 대한 인간의 민주적 통제를 더욱 어렵게 만들 것이고¹²⁾ 또한 LAWs는 특정한 살인에 책임이 있는 사람의 경계를 모호하게 할 수 있다는 점¹³⁾으로 모두 전쟁에 대한 윤리적·도덕적 해이를 초래할 위험성이다.

다) 생명권 경시의 문제

- (1) 치명적 자율무기체계의 진정한 본질은 인류가 창안한 인권에 대한 도전이며, 인권문제는 더 나은 과학과 공학으로도 해결될 수 없는 법철학적 과제이다.
- (2) 따라서 살인 로봇에 대한 가장 본질적이고 핵심적인 쟁점은 기술적인 문제와는 별개의 윤리적 쟁점이라고 할 수 있다.
- (3) 단적으로 치명적 자율무기체계가 합법적인 군사 목표만을 대상으로 한다고 하는 경우에도 기본적인 생명권침해 쟁점을 야기하기 때문이며, 물론 이러한 견해에 대해서는 일상에서는 범죄로 단죄되는 살인은 전쟁에서는 매우 평범하게 허용되는 정당한 행위라는 사실에 비추어 보면 이해가 쉽지 않은 혼란스러운 문제이다.
- (4) 현행 전쟁법의 법리에 입각하는 경우에도 전쟁에 개입되어 있는 전투원이라고 한다면 언제든지 어떤 방법이나 수단에 의해서도 사망할 위험성에 개입되어 있기 때문이다.
- (5) 그러므로 전쟁에 목숨을 내던진 전투원에게 삶에 대한 권리, 즉 생명권이 과연 주어지는 것인가에 대해 의문이 제기되지만, 명백한 것은 인간은 그 누구나 설령 무력충돌 현장에서도 자의적으로는 살해당하지 않는다는 관점에서 본다면 삶에 대한 권리가 그 누구에 의해서도 임의로 부인될 수는 없으며 모든 인간은 비인간적으로 살해되지 않을 권리가 있다.
- (6) 불가피하게 LAWs는 생명권(a right to life) 경시 흐름을 초래할 위험성을 피할 수 없다.

라) 인간의 존엄과 가치의 문제

- (1) LAWs는 단순히 생명권 경시 이상의 인권 쟁점인 인간의 존엄과 가치에 대한 도전을 야기한다.

12) "Deadly Decisions - 8 objections to killer robots", 2016, p.10.

13) Nyagudi, Nyagudi Musandu(2016-12-09)."Doctor of Philosophy Thesis in Military Informatics: Lethal Autonomy of Weapons is Designed and/or Recessive"

- (2) 생명권은 인간의 존엄성(human dignity)에 기초하며, 인간의 존엄성은 단순한 삶에 대한 권리, 즉 생명권보다 더욱더 본질적이며 기본적인 가치이다.
- (3) 인간의 존엄과 가치는 쉽게 포기되거나 방기될 수 없는 인간의 본질적인 존재 이유이며, 이러한 이유로 합법적인 처형인 사형의 경우에도 문명사회에서는 인도적인 절차와 방법을 규정하고 있다.
- (4) 단적으로 존엄과 가치 없는 안전은 시민사회와 국가의 목적을 오해한 것으로. 실제로 킬러-로봇에 대한 비판자들은 인간이 배제된 기계에 자율적 판단에 의한 자동 살인이 인간의 존엄성을 침해한다고 주장한다.¹⁴⁾
- (5) 그렇다면 과연 인간의 존엄성이란 무엇인가? 불행히도 이 개념은 모호하고 일 치된 견해나 이해가 부족하며, 사실 인간사회의 많은 경우에 인간의 존엄성을 결여한 대우가 당연히 불법이 되지는 않는다.
- (6) 거짓말 등은 수치스러운 일로 인간의 존엄과 가치에 배반될 수는 있지만, 형사 처벌에까지는 이르지 않는다.
- (7) 인간의 존엄성은 책임과 구제 그리고 타인 존중과 관련되며, 이러한 가치들은 단순한 기계에는 없는 가치로 로봇-병기에 이러한 가치들의 결여는 단적으로 전술한 ‘의미 있는 인간통제(MHC)’의 부재와 직결된다.
- (8) 의미 있는 인간통제가 없으면 부당한 목표화 오류에 대해 누가 또는 무엇에 대 해 어떤 책임을 져야 하는지가 불분명해질 수밖에 없으며, 당연한 논리의 연장 으로 책임 있는 당사자를 식별할 수 없으면 그 누구도 부당한 살상에 대한 구 제 또는 보상을 요구할 길이 없게 된다.
- (9) 그러므로 인간이 결정해야 하는 가장 심각하고 중대한 결단 가운데 하나가, 인 간을 죽이기 위한 결정을 어떤 경우에도 기계 단독에게 부여하지 않는 것이 인 간 존엄성에 대한 기초적인 고려가 되어야 한다는 논리가 성립된다.
- (10) 그런데 이러한 입장이 합리적인 것처럼 보이지만 예를 들어 책임자를 규정하 거나 지정하는 것만으로 책임 문제를 해결할 수 있는 것은 아니며, 이것은 궁 극적으로 불공정이나 공정 같은 본질적인 정의의 문제를 제기하고 인간의 존 엄과 가치의 엄숙함을 징표 한다고 할 수 있지만 계속해서 남는 의문은 왜 인 간의 도덕적 관심사를 갖지 못하는 기계에 유독 인간 표준을 적용해야 한다는 것인가라는 의문이다.
- (11) 명확한 책임성을 위한 체계를 구축할 수만 있다면 따라서 인간의 존엄성을 보호하 는 데 어느 정도 도움이 될 수 있다면 킬러-로봇이 동원되는 것에 아무런 장애가 없을까?

14) 김자희, 장신, 주성구, 자율 로봇의 잠재적 무기화에 대한 소고 - 개념정립을 통한 규제를 중심으로, 국회 입법조사처, 입법과 정책, 제9권 제3호 2017. 12., pp. 147~148.

마) 인간성의 피폐화와 인간의 수단화 문제

- (1) 불세출의 철학자 임마누엘 칸트는 인간 공동체 사회의 핵심적 도덕법칙으로 2가지 정언명령을 제시했으며, 제2의 정언명령이 인간성 존엄성 목적대우의 원칙이다.¹⁵⁾
- (2) 킬러-로봇은 생명권 경시, 인간 존엄성 와해와 같은 법적 쟁점을 야기하지만 더 나아가 커다란 인권법적 쟁점은 출발부터 사람을 목적으로서가 아니라 단순한 공격대상, 즉 수단으로 취급한다는 사실이다.
- (3) 본질적으로 킬러-로봇은 사람들을 사람들로 인식하지 않고 단순히 대상으로 인식하며, 킬러-로봇에게는 어떤 인간 목표는 공격대상으로 코딩되고, 다른 인간 목표는 죽이지 않는 대상으로 코딩되므로 그 코딩 계정조차도 완전히 정확하지 않을 수는 있지만 그 어떤 경우에도 사람을 목적이 아니라 대상, 즉 수단으로 하여 출발한다.
- (4) 컴퓨터나 로봇은 자연인의 인간이 실제로 하는 것처럼 객체의 본질적인 속성을 인식하지 못하므로 현재 수준에서 인공지능 로봇이 이해하는 것은 약속 기호인 2진법의 1과 0 형태의 숫자 비트 정보로 반인륜적인 인간의 전면적인 수단화는 인간성의 피폐화를 초래한다.
- (5) 인공지능이나 컴퓨터가 인간의 감정을 이해할 수 없기 때문에 치명적인 자율무기 시스템이 충족해야 하는 조건으로 인간 생명과 인격의 존엄성에 대한 존중이 요청되는 것은 당연하다.
- (6) 그러나 사실 전쟁터에서 인간 존중에 대한 요구는 실제로는 존재하지 않는 낭만적인 개념이라는 견해도 있음. 예를 들어 군인들은 방아쇠를 당길 때마다 그 행동의 엄중한 무게감을 고려할 것을 요구하지 않으며, 전쟁터의 군인에게 사살하려는 적군도 희망, 꿈, 두려움을 가진 도덕적 가치 있는 인간으로 대접하며 사살할 것을 요구하는 사람은 거의 없음. 즉, 요구할 수 있는 것은 목표가 합법적인지 아닌지 여부일 뿐이다.
- (7) 전쟁 역사적으로 일부 병사들은 공개적으로 인종 차별적이며 민족주의적이며 증오, 분노 및 기타 퇴폐적인 동기로부터 행동하기도 했으며, 그렇다면 과연 킬러-로봇과 사람은 어떤 차이가 있을까? 사람과 로봇의 차이점은 유능한 사람이 아니더라도 인간 병사에게는 최소한 판단과 인식이 가능한 행위의 올바른 이유 또는 잘못된 이유 또는 정당한 동기부여의 이유를 댈 수 있지만 로봇은 그러한 것들이 당연히 요구되는 상황에서도 절대로 그렇게 하지 못하다는 점이다.
- (8) 따라서 만약 자율무기체계가 인간 수준의 도덕적 존중의 역량이 부족하다면, 인간 수준의 역량에 기초한 인간성 존중의 표준 전투능력을 부여하는 것은 불가능하다는 것이 이론적으로는 제시하고 있다.

15) 한희원, 정의로의 산책(제2판), 삼영사, 2011, pp. 214~216.

- (9) 이에 반하는 조치는 기술의 실패가 아니라 정언 명령적 실패라고 할 수 있으나, 예를 들어 합법적인 총알이나 유도 미사일이 표적을 공격하기에 앞서서 표적에 대해 사려 깊고 존중해야 한다고 요구하는 사람은 아무도 없다.
- (10) 우리는 일정한 의식이 있고 인간과 유사한 인식에 가깝게 행동하는 전투견이나 전쟁에 동원된 다른 동물들에게조차도 행위 전에 그러한 인간적 배려를 요구하지도 않으며, 이는 동물들은 인간에 비해서 도덕적 반성 능력이 태부족하기 때문이다.
- (11) 중요한 것은 이러한 논의나 인식이 없는 LAWs의 무차별적인 등장은 인간성 파괴화를 야기하여 인간을 동물이나 무감각한 로봇으로 하향 수준화시킬 위험성이 킬러-로봇의 윤리문제라고 할 것이다.
- (12) 이러한 이유 등으로 어쨌든 무력충돌에서의 살상에도 도덕적 배려가 반드시 필요하다면 로봇 병기에도 인간이 사전에 일정 한도에서는 프로그래밍할 수 있으며, 그리고 위와 같은 사려 깊고 도덕적인 결정을 내릴 책임은 프로그래머들의 어깨뿐만 아니라 군사행동을 승인하는 정치 지도자들에게도 당연히 부과되어야 한다.

2.2.1.7. 자율살상무기(LAWs)의 윤리적 쟁점 국제법에 따른 자율무기체계에 대한 윤리적 쟁점

가) 국제인도법(군사작전 3대 원칙)

- (1) 국제인도법(IHL¹⁶⁾)의 주요 원칙은 군사적 필요성, 구별성, 비례성, 불필요한 고통의 금지, 공격에 있어 사전주의 원칙으로 국제인도법상 인도주의 원칙 및 마르텐스 조항의 자율무기체계 적용 관련 쟁점 사항이 되두되고 있다.
 - ① 무기체계의 사용으로 민간인 피해를 줄일 수 있다면 오히려 인도주의에 부합하며 사용 가능하다는 입장과 기계의 판단에 의한 인간 생명 박탈 자체가 인도주의에 부합하지 않으므로 허용되서는 안된다는 입장이 대립 중이다.
 - ② 국제인도법은 특정한 무기는 그 무기 사용 자체를 금지하고 있는데, 제네바협약 제1추가 의정서 제36조에서 각 국가들이 연구, 개발, 획득 또는 채택하는 신무기, 전투수단 및 방법의 사용이 일부 또는 모든 상황에서 국제법에 의해 금지되지 않는지 판단하도록 명시하고 있다.

16) 국제인도법(International Humanitarian Law, IHL)은 인도주의적인 이유로 사람과 물건에 대한 무력충돌의 영향력을 제한하기 위한 일련을 규칙들을 말한다. 전쟁법 혹은 무력충돌법이라고도 알려진 국제인도법은 무력충돌 시 전투능력을 상실하였거나 적대행위에 가담하지 아니하는 사람들을 보호하고, 전투의 수단 및 방법을 제한함으로써 무력충돌의 영향력을 최소화하기 위한 국제법이다. 대표적인 국제인도법으로 제네바협약 등이 있다.

- (2) 즉, 국제인도법 상에서 특정한 조건하에서 무기 활용이 제한될 수 있는데, 구별의 원칙(principle of distinction), 비례성의 원칙(principle of proportionality), 사전주의 원칙(principle of precaution)에 부합되어야 한다.
- ① 구별의 원칙: 적법한 공격대상을 적의 군사력을 약화시키는 목표물로 제한하는 것이다. 이에 따라, 전투원과 민간인의 구별, 군사시설 및 물자로부터 민간 물자 및 시설을 구분하는 것이 중요함. 인공지능기술이 고도화되어 이러한 구별에의 활용도가 높아질 가능성이 있는 반면, 외형상 구분이 모호한 경우에 자율무기체계가 과연 구분가능한지가 문제임. 또한, 민간시설이지만 적군이 안에 있는 경우는 원칙들간의 충돌 문제가 있음.
- ✓ 인공지능 LAWs를 전장에 배치했을 때, 신중한 판단을 인공지능 LAWs가 인간수준으로 실행할 수 있는가?
- ② 비례성의 원칙: 군사적 필요성에 의해 얻게 되는 군사적 이익과 공격으로 인해 발생하는 민간인의 부수적 피해를 비교하여 판단하는 것임. 인공지능의 의사결정에 전적으로 의존할 경우 군사적 이익을 위해 대량살상의 가능성이 있으며, 전쟁상황에서의 이익판단은 단순한 산술판단을 넘어선 외교적, 정치적 판단임.
- ✓ 인공지능 LAWs가 군사적 이익과 민간피해 간의 일정한 균형을 유지하는 것은 결코 쉽지 않고, 이를 충족하는 인공지능 LAWs를 구축하기에 상당한 연구와 실험이 필요
- ③ 사전주의 원칙: 자율무기체계와 관련하여 국제법 위반의 문제가 발생할 가능성이 있을 경우 언제나 인간이 개입할 수 있도록 자율무기체계를 설계하고 인간이 통제할 수 없는 상황을 최소화할 것을 요구함.
- (3) 자율살상무기체계 개발과 그의 윤리적 사용에 대한 입장차이가 있으며 그 입장들은 서로 다른 상황 인식과 개념 인식에 기반하고 있다.
- (4) 자율살상무기체계는 기술적인 측면에서 민간인의 피해를 감소시키고 군인의 투입하는 것보다 비용 절감 측면에서 효율적인 반면, 민간인과 전투원의 구분 등이 어려울 수 있으며, 효율성의 이면에는 남용의 가능성과 테러 행위에 사용될 가능성도 제기되고 있다.
- (5) 현재는 국제인권법 등에서 제시된 추상적 원칙만 있을 뿐 이러한 난점을 통제하거나 제어하는 구체적인 실행적 윤리지침이나 준법률적 실천체계가 부재한 상태이다.
- (6) 자율살상무기체계는 무력충돌 상황에서 사용되므로 윤리적 실행 지침 등은 무기의 자율성 정도에 따라 허용되는 것이 아니라 정책 기획, 연구개발, 실험 단계 등의 전 단계를 통해 프로그래밍에 윤리적 설계가 필요하다.

- (7) 이를 통해 인공지능 군사 무기에 대해 의미있는 인간의 통제가 이루어짐으로써 인공지능 무기의 윤리적 사용을 확보할 수 있다.

나) 과잉무기사용금지 원칙

- (1) 군사작전과도 연결되지만 무력충돌 현장에 동원되는 군사무기의 근본적 속성과 그 적법성에 대한 국제인도법(IHL)의 또 다른 중요한 원칙은 과잉피해나 불필요한 고통을 초래하는 무기사용은 금지된다는 원칙이다.
- (2) 이와 같은 과잉 무기 사용금지의 원칙은 핵무기의 적법성에 대한 국제사법재판소(ICJ)의 1996년 자문 의견과 1977년 추가의정서 제35조(2)에 따른 확고한 국제인도법의 기본원칙이지만, 과잉무기사용 금지의 원칙은 공격수단에 대한 것이 아니라 공격의 영향에 강조가 있는 원칙이다.
- (3) 그러므로 무기의 자율성은 본질적으로 과잉무기사용 금지의 원칙이 금지하는 상황을 결정할 때 중요한 요소가 아니라고 할 수 있으며, 과잉무기사용 금지의 원칙은 주어진 군사상황에서 필요이상으로 고통을 초래하는 공격을 금지할 뿐인 것이기 때문이다.
- (4) 물론 자율무기 시스템이 불필요한 손해를 초래할 수 있는 공격의 플랫폼으로 사용될 수도 있지만 그 자체 속성적으로는 반드시 과잉고통을 초래하는 무기는 아니기 때문에 과잉무기사용 금지의 원칙이 LAWs를 전면적으로 금지하는 배지위임장으로 작용되지는 않는다는 주장은 일리가 있어 보인다¹⁷⁾.

다) 무력충돌 원칙 위배책임

- (1) 무력충돌에서 준수해야할 예방의 원칙을 악의적으로 위배하는 행위는 전쟁범죄(war crime)를 구성할 수 있으며, 국제형사재판소에 관한 로마규정(Rome Statute)은 "의도적으로 민간인에 대한 공격을 지시하거나 적대행위에 직접적으로 참여하지 않는 민간인을 공격하는 것"을 전쟁범죄로 규정하고 있다¹⁸⁾.
- (2) LAWs 사용에 대한 국제인도법(IHL)과 국제인권법(IHRL)의 심각한 위반에 대한 형사책임을 포함한 개인 및 국가의 책임 쟁점은 매우 중요한 문제로 철저한 검토와 이해가 필요하다.
 - ① UN 특별보고관 크리스토프 헤인즈(Christof Heyns)가 지적한 것처럼 무력충돌에서 책임성과 억지력 그리고 예방의 약속이 없으면 민간인과 전쟁 범죄피해자의 보호는 낮아질 것이 명백하기 때문이다¹⁹⁾.

17) M. Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A reply to critics', Harv Nat'l Sec J at 2, 2012.

18) Rome Statute of the International Criminal Court, Article 8(War crimes) 2(b).

- ② 따라서 국제인권법(IHRL)과 국제인도법(IHL) 원칙을 위반한 관계자들에 대해 형사책임을 부과하지 않는 것은 그 자체로 별도의 권리 침해로 간주될 수 있으며, 이 경우에 LAWs와 관련하여서는 책임을 져야하는 할 사람들에겐 해당 무기의 소프트웨어 개발자(software developers), 제조업자(manufacturers), 하드웨어 판매자(hardware sellers), 정치인 그리고 군대 지도자가 모두 포함되어야 한다.
- (3) 미국 국방부 지침 3000.09(U.S. Department of Defense Directive 3000.09)는 ‘자동 또는 반자동 무기시스템을 사용하거나 또는 사용을 허가하는 사람은 적절한 주의를 다하여 전쟁법(law of war), 적용 가능한 국제조약, 무기체계 안전규칙 그리고 적용 가능한 제반 규칙을 준수하여야 한다.’라고 규정하고 있다.
- (4) 그러나 무력충돌에 대한 현행 국제법이나 국제관습법에 의하는 경우에도 자율무기시스템에 의해 저질러진 범죄에 대해 제조업체와 소프트웨어 개발자 그리고 군인에 대해 실질적으로 개인책임으로 부과할 수 있는지에 대해서는 불분명하다.
- (5) 완전자율무기의 국제인도법(IHL)과 국제인권법(IHRL)의 중대한 위반에 대해 책임을 개인에게 부담지우는 이론에는 지휘 책임 이론(theory of command responsibility), 국가 책임 이론(theory of state responsibility) 그리고 조직 책임 이론(theory of corporate responsibility) 등이 거론되고 있다²⁰⁾.
- (6) 군대 지휘관이 부하들에 의해 저지른 범죄에 대해서도 명령자 책임을 부담한다는 소위 지휘 책임 이론(command responsibility theory)은 “지휘관이 부하가 어떤 범죄를 저지르려고 계획했음을 알았거나 알 수 있었음에도 불구하고 동범죄를 저지하기 위한 어떠한 조치를 취하지 못했거나 그러한 범죄를 저지른 부하를 사후에도 처벌하지 않은 경우”에 한하여 적용된다는 한계가 있다²¹⁾.
- (7) 자율무기의 경우에 지휘관이 복잡한 인공지능시스템에 의해 저지른 범죄의 가능성을 알고 있었거나 알 수 있었음을 입증한다는 것은 매우 어려운 일이며, 이러한 이유 등으로 LAWs의 개발이나 구매가 개인이 아니라 국가에 의해 주로 이루어지기 때문에 국가에게 책임을 부과하는 것이 보다 현실적이라는 견해가 많다²²⁾.
- (8) 국제 인권단체인 HRW(Human Right Watch)는 ‘마인드 더 갭(Mind the gap)’이라는 제목의 이 보고서에서 현행 법체계가 ‘킬러-로봇’을 개발하고 있는 프로그래머와 제작자, 군 관계자들에게 어떤 책임도 추궁하지 않고 있다고 주장한다.

19) U.S. Department of Defense Directive, No. 3000.09, 21 November 2012, pp. 13~14.

20) Human Rights Watch and the Harvard Law School International Human Rights Clinic, Mind the Gap: The Lack of Accountability for Killer Robots, April 2015.

21) Protocol I additional to the Geneva Conventions, 1977, arts. 86 (2) and 87.

22) D. N. Hammond, Autonomous Weapons and the Problem of State Accountability, Chicago Journal of International Law, Vol. 15(2), 2015.

- ① 보고서는 킬러로봇, 무인자동차 등이 보급되고 있는 가운데 로봇으로 인한 사고가 발생할 경우 어떤 식의 법 적용을 해야 할지에 대해 논의하고 있음²³⁾.
- ② 아무 대가 없이 사람을 죽일 수 있는 ‘킬러-로봇’에 대해 핵심기술이 프로그램머로부터 나오는 만큼 그에 대한 책임을 부담시켜야 한다고 주장함.
- ③ 그러나 또 다른 참가자는 책임 운운하는 주장이 현실을 외면한 무책임한 주장이라고 반발하고 있으며, 책임 추궁보다 또 다른 방안이 필요하다는 주장
- (9) LAWs에 의한 권리 침해가 입증된 경우에는 개별 지휘관이라는 개인보다는 국가에게 책임을 부과함으로써 희생자를 더욱 두텁게 보호할 수 있을 것이라는 사실도 국가책임 원칙이 힘을 얻는 이유이지만 LAWs에 의해 자행된 범죄에 대해 국내법원이 자국의 책임을 인정한다거나 국제법원이 해당 국가의 책임을 인정한다는 판결을 과연 내릴 수 있을 것인지도 쉬운 문제가 아니다.

2.2.2. 자율무기체계의 윤리적 논란 해소를 위한 노력

2.2.2.1. 자율살상무기체계(LAWs) 규제를 위한 국제적 논의

자율살상무기체계의 규제를 공개적으로 논의하는 국제회의가 증가하고 있으며 금지하자는 운동과 캠페인도 전 세계적으로 확산되고 있다. 자율살상무기체계의 개발과 배치, 윤리적 문제를 최초로 제기한 국제기구인 국제 로봇 무기 통제 위원회(ICRAC)²⁴⁾로서 이 위원회는 자율무인전투체계의 연구·개발, 배치, 사용을 금지하는 의제에 대하여 국제적 논의를 제안하였다.

가) 살인 로봇 반대 캠페인

- (1) 자율살상무기체계를 반대하는 또 다른 비정부기구로서 살인 로봇 반대 캠페인(Campaign to Stop Killer Robots)²⁵⁾이 있다.
- (2) 이 캠페인은 지뢰, 섬광 레이저(Blinding Laser), 집속탄(Cluster Munition) 등 무기체계의 사용금지를 촉구해 왔던 이전 운동들처럼 자율살상무기체계를 반대하는 세계적인 운동을 조성하고 이러한 무기체계가 국제법을 위배할 수 있

23) 온라인 포럼인 마이브로드밴드(mybroadband.co.za)에서도 다양한 논의가 진행 중이다.

24) 국제로봇무기통제위원회는 로봇과 연관된 기술, 윤리적 문제, 무기통제, 국제적 관계, 국제인도법 등을 다루고 있는 비영리기구로서 2009년 창설되었다. 이 위원회는 유엔에서 로봇을 무기화함으로써 발생할 수 있는 잠재적 문제들에 대해 논의하였고, 특정재래식무기금지협약을 적용함으로써 로봇의 무기화를 금지해야 한다고 주장하였다.

25) 비정부기구들의 연합체로서 2013년 창설되었다. 이 캠페인은 각 국가와 유엔이 자율살상무기를 개발하는 정책을 무효화하도록 요구하고 있다.

다는 문제점을 제기²⁶⁾하였다.

- (3) 국제적십자위원회(ICRC)도 2014년 자율살상무기체계 전문가회의를 개최하여 현행 국제인도법이 인간에게만 적용되고 기계에 적용은 제한되므로 자율살상무기체계에는 적합하지 않다고 주장하였다.
- (4) 국제인권감시기구(Human Rights Watch)와 국제앰네스티(Amnesty International)는 자율성을 가진 무인전투체계를 전장에서 활용할 때에 국제인도법을 준수하고 윤리적 행동을 위해 선제적 예방조치가 필요하다고 주장²⁷⁾한다.
- (5) 비정부기구와 국제기구가 주도하는 자율살상무기체계의 반대 운동뿐만 아니라 세계적인 전문 지식인들도 자율살상무기체계에 대한 반대 의사를 표명하였다.
- (6) 2015년에 천여 명의 로봇공학자와 인공지능공학자들은 공격용 자율무기체계 자체를 아예 금지하자는 서한을 작성하고 서명하여 공개하였으며,²⁸⁾ 이 그룹에는 스티븐 호킹(Stephen Hawking), 노암 촘스키(Noam Chomsky), 엘론 머스크(Elon Musk) 등 세계적인 유명 학자들도 포함되어 있다.

나) 국제기구 논의

자율살상무기체계를 반대하는 국제사회의 활동이 활발해짐에 따라 유엔은 공식적으로 논의를 시작하였다. 2013년 유엔인권이사회(UN Human Rights Council)에 제출된 특별보고서에는 회원국들의 국제적 합의가 있을 때까지 자율살상무기체계의 개발, 생산, 배치를 중지할 것을 권고하였다. 회의에 참석한 24개국은 자율살상무기체계의 사용에 대해 우려를 표명하였고, CCW가 자율살상무기체계에 적용할 수 있는 적절한 협약이라는 의견도 제시하였다. 이에 따라 2014년 유엔의 CCW 회의에서 자율살상무기체계의 개념과 정의, 기술적 특성, 윤리적·안보적 이슈, 국제 규범화 등이 논의되었으나 이에 대한 구체적인 실행 계획이 도출되지는 않았다.

- (1) 2016년 CCW 회의에서 자율살상무기체계에 대한 논의를 정부 전문가 회의(GGE, Group of Government Experts)²⁹⁾로 전환하자고 협의하여 2017년

26) McVeigh, T. "Killer Robots Must Be Stopped, Say Campaigners."

<http://www.theguardian.com/technology/2013/feb/23/stop-killer-robots>

27) Lewis, M. W. "The Misleading Human Rights Watch and Amnesty International Reports on US Drones."

<http://opiniojuris.org/2013/11/08/misleading-human-rights-watch-amnesty-international-report-s-u-s-drones>

28) Future of Life Institute, "Autonomous Weapons: An Open Letter from AI & Robotics Researches,"

<http://futureoflife.org/open-letter-autonomous-weapons>

29) 정부전문가회의(GGE, Group of Government Experts)는 학계와 민간 사회의 전문가들로부터 특정 이슈에 대해 전문적인 지식과 경험을 바탕으로 토의하고 공유함으로써 국가들이 직면하고 있는 이슈를 해결할 수 있도록 돕는 회의이다.

부터 매년 개최되고 있다.

- (2) 2017년에 GGE에서는 인공지능과 같이 자율살상무기체계에 적용되는 최신 과학기술, 자율살상무기체계 규제의 필요성 등에 대해 논의하였다.
- (3) 2018년 GGE에서는 자율살상무기체계의 개념과 정의에 대한 합의가 도출되지 않았으나 자율살상무기체계의 연구·개발, 획득, 사용의 국제인도법 합치 의무, 기계가 아닌 인간에게 책임 귀속, 새로운 자율무기체계 개발 시 위험 평가와 완화 조치의 필요성 등 11개의 원칙³⁰⁾이 수립되었다.
- (4) 2019년 유엔의 CCW 회의에서는 많은 국가들이 자율살상무기체계를 금지하거나 규제하는 새로운 국제법의 개발에 지지를 보냈으며, 동년도 GGE에서는 자율살상무기체계의 국제인도법 도전과 군사적 응용 가능성, 인간-기계의 상호작용 등의 의제에 대해 논의가 이루어졌다.
- (5) 2020년 3월 LAWs가 이미 리비아에서 사용되었으며, 해당 무기체계는 운용자와 무기 사이 데이터 연결 없이 목표물을 공격하도록 프로그래밍 되었다고 한다. 이러한 상황에서 ‘킬러-로봇 반대 운동(Campaign to Stop Killer Robots)’과 같은 국제기구들과 많은 국가들은 ‘규범 창출자(norm entrepreneurs)’로서 LAWs를 국제협약으로 금지하기 위해 규범화를 시도하고 있다.

2.2.2.2. 자율살상무기체계(LAWs) 활용에 대한 주요 국가별 견해와 입장

가) 유엔 회원국들의 입장

- (1) 2019년 유엔의 CCW 회의에서 아르헨티나, 오스트리아, 브라질 등 28개국은 자율살상무기체계의 개발, 보유, 사용에 대한 완전한 금지를 주장했고 다른 일부 국가들은 자율살상무기체계에 대해 인간의 통제를 보장하는 법적 합의가 필요하다고 주장³¹⁾하였다.
- (2) 또 다른 국가들은 자율살상무기체계의 사용에 대해 일정 수준의 통제는 필요하나 선부른 규제는 너무 앞서 나가는 것이라고 강조하고 있다.

30) 2018년 GGE에서 제시된 자율살상무기체계에 대한 세부 이행 원칙은 다음과 같다: ① 모든 무기체계에 대한 국제 인도법의 적용, ② 무기체계 사용 관련 인간의 책임 유지, ③ 신종 무기개발, 배치 및 사용 관련 국제법에 따른 책임성 확보, ④ 신무기 도입 시 국제법상 금지 여부 결정, ⑤ 자율살상무기 분야 기술에 기반한 새로운 무기의 개발 및 획득 시 적절한 안전조치 및 비확산조치 고려, ⑥ 모든 무기체계에 있어 위험감소 조치 확보, ⑦ 국제인도법 준수에 있어 자율살상무기 분야 신기술 사용 고려, ⑧ 자율살상무기 분야 신기술의 인격화(anthropomorphize) 금지, ⑨ 자율기술의 평화적 사용 방해 금지, ⑩ 군사적 필요성과 인도적 고려간 균형을 추구하는 특정재래식무기금지협약이 동 문제 논의 틀로서 적절함.

31) Campaign to Stop Killer Robots. "High-Level Concerns on Killer Robots at UN." <http://www.stopkillerrobots.org/> 2019/10/unga74

나) 주요 국가별 입장

(1) 미국

- ① 유엔을 비롯한 국제 사회에서 자율살상무기체계에 대해 공식적 활동을 가장 왕성하게 추진하는 대표적인 국가는 미국이다.
- ② 미국은 국제적 논의 방향을 미국 정책에 부합하게 이끌어가려는 의지를 강하게 보이고 있는데, 자율무기체계에 대한 미국의 핵심적 이슈는 인간의 통제 수준이 아니라, 자율살상무기체계가 지휘관의 의도를 얼마나 구현할 수 있는지와 인간으로서 적절한 판단을 얼마만큼 보조할 수 있는지에 있다고 주장한다.
- ③ 또한 자율살상무기체계의 사용에 대한 국제적 금지를 반대하고 있으며 현 기술 수준이 규제할 단계는 아니라는 입장을 내세우고 있다.
- ④ 미국과 유사한 입장을 보이는 국가들로서 러시아, 호주와 이스라엘을 들 수 있는데 이 국가들은 자율무기체계에 대한 장점을 강조하며 법적으로 추가적인 규제를 만드는 것에 대해 부적절하다는 입장이다.

(2) 중국

- ① 이와는 다르게 중국은 자율살상무기체계에 교전권과 살상권을 부여하며, 자율살상무기체계의 의사결정 과정에 인간의 개입을 최소화하여 결정한 행동을 인간이 종료시킬 권한이 없다고 본다.
- ② 즉, 자율살상무기체계에 부여된 자율권은 완전한 자율권을 의미하며, 이를 미래 전쟁에 활용하는 데에 긍정적인 입장이라 적극적으로 개발하고 있다³²⁾.

(3) 러시아

- ① 자율살상무기체계가 중요 시설에 대한 방호, 주요 표적 공격, 테러리스트 제거, 민간인 보호 등의 임무 영역에서는 사용될 수 있으므로 자율살상무기체계의 활용에 매우 긍정적으로 찬성하고 있다³³⁾.
- ② 가장 최근인 2018년 GGE에서는 이 국가들을 포함해서 총 12개국이 자율살상무기체계에 대해 국제적 조약을 만들어 타결하는 것에 반대 의사를 표시하였다.

(4) 기타 국가

- ① 오스트리아, 브라질, 칠레 등 일부 국가에서는 자율살상무기체계에 대한 인간의 통제가 자율성 수준과 불가분의 관계에 있기 때문에 통제가 전제되어야 한다고 주장한다.

32) United Nations Geneva. "Position Paper, Submitted by China," CCW/GGE.1/2018/WP.7, para.3.

33) United Nations Geneva. "Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapon Systems in the Context of the Purposes and Objectives of the Convention, Submitted by the Russian Federation," CCW/GGE.1/2018/WP.6, para.2(a).

- ② 또한 GGE가 자율살상무기체계에 대해 법적 구속력이 있는 제도를 만들어야 한다는 입장을 보이고 있다.
- ③ 프랑스와 독일은 위와 같은 주장들에 대해 중간 정도의 위치에 있으며, 이 국가들은 자율살상무기체계의 자율성이 의사결정 프로세스를 향상시킬 수 있고 자율살상무기체계의 운영 결과에 대해서는 운용자의 책임의무가 있다고 강조한다.
- ④ 따라서 자율살상무기체계에 대한 인간의 통제와 책임에 대한 가이드 라인을 만들고 국제적인 합의를 이루어야 한다는 입장이다.
- ⑤ 만약 CCW에서 국제적 합의가 이루어지지 않는다면 규제보다는 국가별 정치적 선언 같은 방식이라도 추진해야 한다고 주장하고 있다.

2.2.2.3. 자율살상무기체계(LAWs) 개발·사용·생산에 대한 국제사회의 동향 및 찬·반론

가) 국제사회의 동향

- (1) LAWs는 일반인들에게도 『킬러-로봇』이 지구 전역을 배회하는 위험성에 대한 중대한 논쟁을 불러일으켰으며, 이에 2013년 출범한 킬러-로봇 금지 국제캠페인(Campaign to Stop Killer Robots³⁴⁾)은 『자율무기체계 개발·생산·사용에 대한 포괄적이고 선제적 금지』를 구호로 내세우면서 국제법의 제정과 각국의 국내법 제정을 포함한 관련 조치를 국제 사회에 호소하고 있다³⁵⁾.
- (2) 국제인권 NGO인 휴먼 라이츠 워치는『LAWs 개발, 생산 및 사용에 대한 선제적 금지』를 요구하면서 LAWs가 무력 충돌에서의 구별의 원칙과 비례의 원칙 그리고 군사적 필요성의 원칙을 포함하여 국제인도법의 표준 요구 조건을 충족할 수 있는지에 대해 의문을 제기했다³⁶⁾.
- (3) 2014년 국제여성평화·자유연맹(Women's International League for Peace and Freedom: WILPF)은 LAWs가 의미 있는 인간 통제 없이 작동할 수 있는 가능성에 심각한 우려를 제기하고 자율살상무기체계를 프로그래밍하는 자체가 이성과 법에 대한 근본적인 도전이라고 LAWs 그 자체를 비판했다³⁷⁾.
- (4) 국제 적십자위원회(ICRC)는 자율무기체계의 개발중단이나 생산 및 사용금지 같은

34) 2012년 창설된 국제 NGO. <https://www.stopkillerrobots.org/>

35) Campaign to Stop Killer Robots, <http://www.stopkillerrobots.org/the-solution/>

36) Human Rights Watch, <http://www.hrw.org/topic/arms/killer-robots>.

37) Women's International League for Peace and Freedom(WILPF), Statement to the CCW Meeting of Experts on Lethal Autonomous Weapon Systems, 2014,

국제운동에는 참여하지 않기로 선택하면서도, 각국은 LAWs의 진전된 개발이나 무력 현장에의 배치에 앞서서 “자율무기의 핵심적 기능(critical functions)에 있어서 무기 자체의 자율성이 제기하는 근본적인 법적 및 윤리적 문제를 고려할 것”을 촉구했다.

- (5) ICRC는 또한 자율무기의 핵심적 기능을 포함하여 각국이 개발하거나 구입하는 신무기와 전쟁의 신기술에 대한 완전한 법적 재검토를 할 것을 촉구했다³⁸⁾.

나) 반대론

- (1) 2015년 7월 1,000명이 넘는 인공지능 전문가들이 인공지능 자율무기 금지를 촉구하는 결의안에 서명했으며, 결의안은 부에노스아이레스에서 개최된 제24회 인공지능 국제합동회의(IJCAI-15)에 제출되어 대표적으로 스티븐 호킹(Stephen Hawking), 엘론 머스크(Elon Musk), 스티브 워즈니악(Steve Wozniak), 노암 촘스키(Noam Chomsky), 스카이프(Skype)의 공동 창립자인 잔 탈린(Jaan Tallinn), 구글 딥마인드(Google DeepMind) 공동 발명자인 데미스 허사비스(Demis Hassabis) 등이 추가 서명했다.
- (2) 서한은 “치명적인 자동화 무기는 일단 개발되면 더 큰 규모로 무장 충돌을 일으킬 것으로 전쟁의 3차 혁명이 될 수 있다며 인류의 피해가 엄청난 수준에 이를 것이고 킬러-로봇이 전투의 문턱을 낮춰 인간의 삶을 크게 훼손할 것”이라고 주장
- (3) 한편 2016년 버락 오바마 전 미국 대통령은 재임 초기에 중에 드론을 전쟁에 동원하는 것에 흔들렸고 만약 어떤 미국 대통령이 드론을 포함함 자율무기를 전쟁에 동원한다면 이제 전쟁은 전 세계를 통틀어 영원한 전쟁 그리고 어떠한 법적 책임이나 민주적 논의도 없는 수많은 비밀 공작적 전쟁이 될 것이라고 경고하고 있다³⁹⁾.

다) 찬성론

- (1) 찬성론자들은 인류사에서 전쟁을 피할 수 없다면 차라리 로봇을 활용해 방위비 부담을 줄이고 인명 손실을 최소화하자는 의견이다.
- (2) 남호주대학 산하 방위시스템연구소 소장인 안토니 핀 교수는 “자율살상무기는 ‘파이어 앤 포켓(Fire & Forget)’ 무기로서 재래식 무기보다 민간인 피해를 절대적으로 줄이고 전쟁을 조기에 끝낼 수 있을 것”이라고 주장한다.

38) ICRC, supra note 04.

39) “Barack Obama on 5 Days That Shaped His Presidency”. Daily Intelligencer.

- (3) 로봇 개발에 찬성하는 측의 윤리적인 설득력도 만만치 않음. 즉, 꼭 필요한 명령만 내려 로봇 무기가 무고한 살상을 방지할 수 있는 프로그램을 개발하면 예컨대 로봇은 여성들을 성폭행하지 않는 것처럼 불필요한 피해를 최소화할 수 있는 장점도 있다고 주장한다.
- (4) 또한 LAWs 반대론자들은 LAWs에 의한 살인 책임 책임자의 경계가 모호하게 될 수 있다는 점을 우려하지만 Thomas Simpson과 Vincent Muller는 오히려 누가 명령을 내렸는지 기록하기가 더 쉬워 책임자도 명확하게 가려낼 수 있다고 주장한다⁴⁰⁾.

라) 주권 국가의 입장

- (1) 중국이나 러시아 같은 사회주의 국가나 북한이나 이란 같은 통제국가의 치명적 자율무기 연구와 개발은 과학기술의 문제일 뿐으로 법적·윤리적 문제를 배제해 온 것이 무기개발의 역사였다.
- (2) 영국 국방부는 2011년 “현재까지는 인간의 개입 없이 작동하는 자율무기 시스템을 개발할 의향이 없지만 치명적 자율무기가 국방 시스템을 보다 효과적으로 만드는 자동화 수준을 높이는데 기여할 수 있을 것”으로 본다고 말했다⁴¹⁾.
 - ① 치명적인 자율무기를 개발하지 않았고 영국 군대에서 쓰이는 모든 무기는 인간의 감시와 통제하에 있을 것이라면서 “고도로 자동화된 시스템보다는 원격 조종 시스템 개발 노력에 집중하고 있다”라고 덧붙였다.
 - ② 2015년에는 오히려 치명적인 자율무기에 대한 금지 조치에 반대하며 “국제 관련 법은 이미 이 분야에 충분한 규제를 제공하고 있다”라고 밝혔다.
- (3) 미국은 ‘자율무기에서도 인간 통제력’을 절대 놓아선 안 된다는 입장으로 전장 상태를 판단하는 건 인공지능에 맡기되 마지막 공격 스위치는 인간의 몫으로 남겨뒀다고 주장한다.
- (4) 이에 미국 국방부는 군사용 로봇을 100% 인공지능 로봇이 아닌 『지능 확장(IA·Intelligence Augmentation)형 로봇』 개념으로 개발할 것이라면서 미국의 정책은 『자율무기 체계는 지휘관과 운영자가 자율무기 사용에 대한 적절한 수준의 직접적인 인간 판단을 할 수 있도록 설계되어야 한다』는 원칙이라고 주장했지만, 미국도 사람을 살상하거나 인간의 개입 없이 공격 목표를 설정하고 자율적으로 공격하는 자체적으로 설정한 소위 적절한 수준(appropriate levels)

40) Simpson, Thomas W; Muller, Vincent C. (2016). “Just war and robots' killings”. Philosophical Quarterly.

41) Ministry of Defence of the UK, Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems, 30. 3. 2011.

- 과 기준을 충족하는 자율무기 시스템을 이미 인증했다⁴²⁾.
- (5) 미국은 또한 LAWs가 '떠오르는 새로운 기술'로서 원격 조종 항공기나 기존의 다른 무기 체계의 범주에 속하지 않는다고 인식하고 있다⁴³⁾.
 - (6) 2015년 미국 해군은 해군의 무인무기시스템 책임자를 임명하면서 미래에는 무인무기시스템이 『새로운 정상(new normal)』이 되어야 한다고 말한다⁴⁴⁾.
 - (7) 로버트 워크(Robert Work) 국방부 부장관은 2016년 『국방부는 치명적 권한(lethal authority)을 기계에 위임하지 않을 것이지만 어떤 독재 정권이 그렇게 할 수 있기 때문에 이것을 다시 생각할 필요가 있다.』고 말한다⁴⁵⁾.
 - (8) 한편 멕시코, 칠레, 쿠바, 이집트, 교황청, 파키스탄 등은 자율무기에 대한 선제적이고 포괄적인 금지를 요구했으며, 반면에 일본, 스리랑카, 스위스, 프랑스, 아일랜드, 대한민국은 LAWs를 반대하지 않으면서 『의미 있는 인간통제(MHC)』 개념 토론에 참여할 용의를 밝혔다⁴⁶⁾.

2.2.2.4. 자율살상무기(LAWs) 윤리적 논란 해소방안

- 가) 기존 제안된 국내외 인공지능윤리 권고안들은 추상적 원칙들로 이루어져 있어 구체적인 실행 방안이 필요하다.
- 나) 기술발전에 따라 해결 가능한 윤리적 쟁점과 해결 불가능한 윤리적 쟁점을 구분하여 추상적인 쟁점 논의를 벗어나 현실적인 논의가 필요하다.
- 다) 미 국방부가 2020년 2월 발표한 인공지능에 대한 윤리원칙이나 그를 시스템에 적용하는 가이드 라인이 제시되어 군사 인공지능 관련 개발 기업에 적용되는 것처럼, 한국에서도 면밀한 검토를 통해 한국 군사 상황에 맞는 군사 인공지능 윤리 개발 가이드 라인이 필요하다.
- 라) 해결 가능한 쟁점들은 윤리적인 설계 및 검수를 거친 인공지능 군사 무기가 인간의 통제 하에 있을 때 오히려 전시에서의 인명 훼손을 이전보다 감소시킬 수 있으므로, 인공지능 무기의 기획, 설계, 개발 등의 전체 단계를 통해

42) US Department of Defense(2012). "Directive 3000.09, Autonomy in weapon systems" p. 2; Gubrud, Mark (April 2015). "Semi-autonomous and on their own: Killer robots in Plato's Cave". Bulletin of the Atomic Scientists. 더 나아가 자율적으로 목표물을 식별하고 공격하는 반 자율적 헌터킬러는 인증조차 필요로 하지 않는다.

43) U.S. Delegation Opening Statement as delivered by M. Meier, CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 13 April 2015.

44) J. Sydney, Jr. Freedberg, Secretary Of Drones: Mabus Creates DASN for Unmanned, Breaking Defence, 2015.

45) Pentagon examining the 'killer robot' threat". Boston Globe. 30 March 2016.

46) Campaign to Stop Killer Robots, Country Views on Killer Robots, 14 April, 2016.

윤리적 점검사항들을 정하여 적용할 필요가 있다(인공지능윤리 자율점검 항목들로 구현될 필요가 있음).

- 마) 인공지능 발전의 현 단계에서 해결이 어려운 쟁점들은 자율살상무기가 완전한 자율시스템의 수준이 될 때 피아식별이나 전투원-민간인 구분에 있어 인간보다 훌륭한 성능을 보일 수 있는지의 여부이며 이는 인공지능기술의 발전에 달려있다.
- 바) 자율무기체계의 의사결정으로 인간에게 불필요한 피해를 발생시키는 가능성이 제기되지만 이는 인간의 제어가 불필요한 가장 높은 단계의 자율성을 가지는 인공지능 수준에서 제기될 문제이며, 이에 앞서 인간의 제어가 필수적인 반(semi)자율 인공지능에서의 윤리적 설계와 인간의 역할을 구체화할 필요성이 있다.
- 사) 강대국들은 군사무기 관련 인공지능윤리 규범을 타국을 제어할 수 있을 때 적극 찬성하고 자국에 이익이 되는 한에서만 찬성한다.
- 아) 따라서 외국의 군사무기 윤리 권고를 그대로 받아들이기보다는 현 한국 군사 관련 현실적 상황 맥락을 검토하여 한국의 이익을 고려하여 수정 및 재구성하는 것이 필요하다.
- 자) 인공지능 편향은 반드시 부정적인 것이 아니라 특정 목적의 인공지능에 있어서는 필수적인 고려 요소이므로 군사 무기라는 특정 목적을 가지는 자율무기체계에 있어서는 데이터 및 알고리즘 편향이 인공지능시스템 설계 과정에서 자국의 이익을 침해하는 부분과 이익에 도움이 되는 부분을 분리할 필요성이 있다.
- 차) 이는 전시상황에서 구분하지 않아도 되는 특성들과 반드시 구분해야 하는 특성들을 정리함으로써 가능해지며 각 국가의 적국의 군사 특성마다 달라지므로 한국의 자국과 적국의 정의 및 그 특성에 따라 구성될 필요가 있다.

2.3. 소결론

2.3.1. 일반적인 개념의 윤리

2.3.1.1. 일반적인 윤리의 개념

- 가) 윤리적 원칙들은 상호 연관성이 있으며, 사회적 조화와 공정한 교류를 이끌어내기 위한 기초 형성 또는 사회적 동의를 위한 거버넌스 구축을 추진하고 있다.

- 나) 이를 개인, 조직, 사회 수준에서 적용하여 삶의 다양한 영역에서 윤리적인 행동을 실천하는 것이 중요하다.

2.3.1.2. 윤리와 도덕의 관계

- 가) 자유주의의 윤리적 개념과 공동체주의의 도덕적 개념으로 접근할 수 있으며, 일반적인 윤리의 기준은 자유주의의 윤리적 개념과 공동체주의의 도덕적 개념을 반영하여 사회 구성원들의 합의에 의해 형성된다고 할 수 있다.
- 나) 즉, 일반적인 윤리는 사회적으로 공유되고 인정된 규범과 가치를 기반으로 한 행동의 올바름과 옳음에 대한 지침을 나타내는 것이며, 이는 개인, 집단, 조직 또는 사회 전체의 행동 및 결정에 영향을 미치고 상호 존중, 정의, 성실, 책임, 배려, 정직 등의 원칙에 근거한다.

2.3.2. 군사적 측면에서의 윤리

2.3.2.1. 군사용 인공지능의 주요 특성

- 가) 군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.
- 나) 군사용 인공지능 자율무기체계의 개념 및 범위는 군사 로봇에 포함된 자동화 무기(Autonomous Weapons)를 의미하며, 이를 통틀어 치명적인 자동화 무기체계(LAWs, Lethal Autonomous Weapons Systems) 혹은 자율살상 무기라 부르기도 한다.
- 다) 군사용 인공지능 자율무기체계의 유형은 자동적 방어체계와 자동적 공격체계로 분류할 수 있으며, 인간이 배제된 신속한 대응을 할 수 있는 체계이다.
- 라) 군사용 인공지능 중 자율무기체계는 물리적 교전, 전자 공격, 사이버 방호 등의 다양한 임무에 활용되며 인간 운용자의 관제를 받는 자율체계까지 포함되며, 이 중 자율살상무기체계의 구체적인 정의에 대해서는 국제적으로 합의된 바 없으나 자율살상무기는 자율성과 교전권이 부여된 특정 자율무기체계를 나타내는 포괄적인 개념으로 정의할 수 있다.
- 마) 자율무기체계의 윤리적 쟁점은 『위협의 크기, 특성, 양상의 변화 - 기존 위협의 확대, 위협의 특성 변화, 새로운 위협의 출현』과 『기계에 의한 의사결

정 및 살인의 책임소재가 모호하다는 점, 『전쟁에 대한 의사결정이 추상적이 될 위협』 측면에서 대두되고 있다.

- 바) 특히 자율살상무기(LAWs)의 윤리적 쟁점은 『법적·윤리적 측면』, 『윤리적·도덕적 해이의 문제』, 『생명권 경시의 문제』, 『인간의 존엄과 가치의 문제』, 『인간성의 피폐화와 인간의 수단화 문제』 등으로 대두되고 있다.
- 사) 국제법에 따른 자율무기체계에 대한 윤리적 쟁점은 『국제인도법(군사작전 3대 원칙-구별의 원칙, 비례성의 원칙, 사전주의 원칙)』, 『과잉무기사용금지 원칙』, 『무력충동 원칙 위배 책임』 등으로 대두되고 있다.

2.3.2.2. 군사적 측면에서 윤리적 논란

가) 자율살상무기체계 규제를 위한 국제적 논의

- (1) 자율살상무기체계의 개발과 배치, 윤리적 문제를 최초로 제기한 국제기구인 국제 로봇 무기 통제 위원회(ICRAC)로 자율무인전투체계의 연구·개발, 배치, 사용을 금지하는 의제에 대하여 국제적 논의를 제안
- (2) 비정부기구로서 살인 로봇 반대 캠페인(Campaign to Stop Killer Robots)과 국제적십자위원회(ICRC)에서 국제인도법에 위배될 수 있다는 문제점을 제기
- (3) 또한, 국제 인권 감시기구(Human Rights Watch)와 국제 앰네스티(Amnesty International)도 전장에서 활용할 때에 국제인도법을 준수하고 윤리적 행동을 위해 선제적 예방조치가 필요하다고 주장
- (4) 유엔은 2013년 유엔 인권 이사회(UN Human Rights Council)에 제출된 특별보고서에는 회원국들의 국제적 합의가 있을 때까지 자율살상무기체계의 개발, 생산, 배치를 중지할 것을 권고
- (5) 2014년 유엔의 CCW 회의에서 자율살상무기체계의 개념과 정의, 기술적 특성, 윤리적·안보적 이슈, 국제 규범화 등이 논의되었으나 이에 대한 구체적인 실행 계획이 도출되지 못함

나) 자율살상무기(LAWs) 개발·사용·생산에 대한 국제사회의 동향 및 찬·반론

- (1) 찬성론은 방위비 부담을 줄이고 인명 손실을 최소화하자는 의견, 윤리적 측면에서 꼭 필요한 명령만 내려 로봇 무기가 무고한 살상을 방지할 수 있는 프로그램 개발하면 피해를 최소화할 수 있는 장점, LAWs에 의한 살인 책임 책임자의 경계는 누가 명령을 내렸는지 기록하기가 더 쉬워 책임자도 명확하게 가려낼 수 있다고 주장

- (2) 반대론은 킬러-로봇이 전투의 문턱을 낮춰 인간의 삶을 크게 훼손할 것이라는 부분, 전 세계적으로 영원한 전쟁 그리고 어떠한 법적 책임이나 민주적 논의도 없는 수많은 비밀 공작 전쟁이 될 것이라고 경고, 근본적인 법적 및 윤리적 문제(이성↔법), 국제법 위배 등으로 반대

다) 자율살상무기체계 활용에 대한 주요 국가별 견해와 입장은 <표 2-1>과 같음.

<표 2-1> LAWs 활용에 대한 주요 국가별 견해와 입장

구분	미국	중국	러시아	기타국가
자율살상무기 체계 도입/사용	• 도입 필요 • 사용금지 반대	• 도입 필요 • 사용금지 반대	• 도입 필요 • 사용금지 반대	• 아르헨티나, 오스트리아, 브라질 등은 사용금지 찬성
자율권 부여, 인간 개입/통제	• 일부 자율권 부여 • 교전권, 살상권은 통제 • 일정 수준의 인간 개입 필요	• 완전 자율권 부여 • 교전권, 살상권은 부여 가능 • 인간 개입 최소화	• 완전 자율권 부여	• 영국, 프랑스, 독일은 미국과 유사한 입장
국제적 규제의 필요성	• 국제적 규제에 반대	• 국제적 규제에 반대	• 국제적 규제에 반대	• 남미, 아프리카 등 비동맹국가 위주로 국제적 규제 필요성 주장

2.3.2.3. 윤리적 논란 해소방안

- 가) 기존 제안된 인공지능 윤리 권고안들은 추상적 원칙들로 이루어져 있어 구체적인 실행 방안이 필요
- 나) 미 국방부가 2020년 2월 발표한 인공지능에 대한 윤리원칙 및 적용하는 가이드 라인이 제시, 한국 군에서도 면밀한 검토를 통해 한국 군사 상황에 맞는 군사 인공지능 윤리 개발 가이드 라인이 필요
- 다) 실질적인 해소 방안으로 윤리적인 설계 및 검수를 거친 군사용 인공지능은 인간의 통제하에 있을 때 오히려 전에서의 인명 훼손을 이전보다 감소시

킬 수 있으므로, 인공지능 무기의 기획, 설계, 개발 등의 전체 단계를 통해 윤리적 점검사항들을 정하여 적용할 필요(인공지능 윤리 자율점검 항목들로 구현될 필요가 있음).

- 라) 자율살상무기체계가 완전한 자율시스템의 수준이 될 때 피아식별이나 전투원-민간인 구분에 있어 인간보다 훌륭한 성능을 보일 수 있는지에 대한 문제임.
- 마) 가장 높은 단계의 자율성을 가지는 인공지능 수준에서 제기될 문제로 이에 앞서 인간의 제어가 필수적인 반(semi)자율 인공지능에서의 윤리적 설계와 인간의 역할을 구체화할 필요성이 있음.
- 바) 현 한국 군사 관련 현실적 상황 맥락을 검토하여 한국의 이익을 고려하여 수정 및 재구성하는 것이 필요
- 사) 인공지능 편향은 반드시 부정적인 것이 아니며 군사용 무기라는 특정 목적을 가지는 자율무기체계에 있어서는 데이터 및 알고리즘 편향이 인공지능시스템 설계 과정에서 자국의 이익을 침해하는 부분과 이익에 도움이 되는 부분을 분리할 필요성이 있으며, 이는 군사적 특성마다 달라지므로 한국의 자국과 적국의 정의 및 그 특성에 따라 구성될 필요가 있음.

3. 인공지능 윤리 이론 및 사례

3.1. 인공지능 윤리 관련 이론

3.1.1. 인공지능이란?

3.1.1.1. 정의

가) 인공지능은 인간의 지적 능력을 기계로 구현하는 과학기술로서, ‘인간의 학습능력과 추론능력, 지각능력, 자연언어의 이해능력 등 인간의 지적 능력을 컴퓨터 프로그램으로 구현하는 과학기술’로 정의할 수 있다.

나) 존 매카시(John McCarthy)는 1956년 다트머스회의(Dartmouth Conference)에서 ‘인공지능’이라는 말을 처음 사용하였으며, 인공지능을 『지능적인 기계를 만드는 과학 및 공학(The science and engineering of making intelligent machines)』로 정의하고, 인공지능의 궁극적인 목표 : 사람과 유사한 혹은 사람 수준의 기계장치를 개발하는 것이라고 했다.

- ✓ 실제로는 인공지능(Artificial Intelligence)이라는 용어는 다트머스회의 개최를 위해 1955년 8월 31일 작성된 제안서(A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence)에 최초로 등장

인공지능(AI: Artificial Intelligence)이란?

인간의 지적능력을 컴퓨터로 구현하는 과학기술로서, ①상황을 인지하고, ②이성적·논리적으로 판단·행동하며, ③감성적·창의적인 기능을 수행하는 능력까지 포함



〈그림 3-1〉 인공지능이란?(인공지능 국가전략, 관계부처합동)

3.1.1.2. 인공지능의 이점

가) 복잡한 문제 해결

- (1) 인공지능 기술은 ML 및 딥-러닝 네트워크를 사용하여 인간과 유사한 지능으로 복잡한 문제를 해결
- (2) 인공지능은 패턴을 발견하고, 정보를 식별하고, 해답을 제시하는 등 대규모로 정보를 처리
- (3) 인공지능을 사용하면 사기 탐지, 의료 진단, 비즈니스 분석과 같은 다양한 분야의 문제를 해결

나) 비즈니스 효율성 증대

- (1) 인간과 달리 인공지능 기술은 성능 저하 없이 연중무휴로 작동할 수 있다. 즉, 인공지능은 수동 작업을 오류 없이 수행
- (2) 인공지능으로 반복적이고 지루한 태스크를 집중적으로 처리하면 비즈니스의 다른 영역에 인적 자원을 사용
- (3) 인공지능을 사용하면 모든 비즈니스 관련 태스크를 간소화하는 동시에 직원의 업무량을 줄일 수 있음.

다) 스마트한 의사결정

- (1) 인공지능은 ML을 사용하여 인간보다 더 빠르게 대량의 데이터를 분석할 수 있음.
- (2) 인공지능 플랫폼은 추세를 파악하고, 데이터를 분석하고, 지침을 제공할 수 있음.
- (3) 인공지능의 데이터 예측은 향후 조치를 위한 최선의 방법을 제안하는 데 도움이 될 수 있음.

라) 비즈니스 프로세스 자동화

- (1) ML로 인공지능을 훈련하여 태스크를 정확하고 빠르게 수행할 수 있으며, 이렇게 하면 어렵거나 지루한 비즈니스 작업을 자동화하여 운영 효율성을 높일 수 있음.
- (2) 인공지능 자동화를 사용하여 더 복잡하고 창의적인 작업에 필요한 직원 리소스를 확보할 수 있음.

3.1.1.3. 신뢰도 높은 인공지능 구현을 위한 당면 과제

가) 데이터 거버넌스

- (1) 데이터 거버넌스 정책은 규제 제한 및 개인 정보 보호법을 준수해야 함.
- (2) 인공지능을 구현하려면 데이터 품질, 개인 정보 보호, 보안을 관리해야 함.
- (3) 고객 데이터 및 개인 정보 보호에 대한 책임은 해당 조직에 있으며, 데이터 보안을 관리하려면 각 계층에서 인공지능 모델이 고객 데이터를 어떻게 사용하고 상호 작용하는지 명확하게 이해해야 함.

나) 기술적 어려움

- (1) 기계 학습으로 인공지능을 훈련하려면 막대한 리소스가 필요하며, 딥-러닝 기술이 제대로 작동하려면 높은 처리 파워 임계값이 필수적임.
- (2) 인공지능 애플리케이션을 실행하고 모델을 훈련하려면 강력한 컴퓨팅 인프라가 있어야 함.
- (3) 처리 파워는 비용이 높기 때문에 인공지능 시스템의 확장성을 제한할 수 있음.

다) 데이터 제한

- (1) 편향 없는 인공지능 시스템을 훈련하려면 엄청난 양의 데이터를 입력과 훈련 데이터를 다루고 처리할 수 있는 충분한 스토리지 용량
- (2) 훈련에 사용하는 데이터의 정확성을 보장하기 위해 효과적인 관리 및 데이터 품질 프로세스를 마련

3.1.2. 인공지능 윤리에서 말하는 윤리의 개념

3.1.2.1. 개요

인공지능 윤리란 인공지능 기술이 발전하면서 발생할 수 있는 윤리적 문제를 다루는 분야로 인공지능은 인간의 생활과 밀접하게 연관되어 있으며, 다양한 분야에서 활용되고 있다. 하지만 인공지능이 인간의 권리와 자유를 침해하거나 잘못된 판단을 내리는 등의 문제가 발생할 수 있다. 이에 인공지능 윤리는 인공지능 기술 개발과 사용에 있어서 인간의 가치와 존엄성을 존중하고, 투명성과 책임성을 강화하여 안전하게 활용될 수 있도록 하는 것이 중요하다. 이를 위해 국제적인 기구와 단체에서는 인공지능 윤리 가이드 라인을 제정하고 있으며, 기업과 연구기관에서도 자체적으로 윤리 지침을 마련하여 준수하고 있다.

즉, 인공지능에서 말하는 윤리는 인공지능 시스템이 가지고 있어야 하는 윤리적 원칙, 가치, 행동 지침을 의미하는 것으로 인공지능은 점점 더 사회 시스템에서 중요한 역할을 수행하며 사회, 경제, 문화, 정치 등 다양한 영역에 영향을 미치게 되면서 그 사용과 개발에 대한 윤리적 고려가 더욱 중요해지고 있는 실정이다.

3.1.2.2. 인공지능의 윤리적 원칙

- 가) 공정성과 투명성: 알고리즘의 공정성과 의사결정 과정의 투명성을 보장하여, 편견과 차별을 방지하고 공정한 의사결정을 유도한다.
 - 나) 개인정보 보호와 프라이버시: 사용자의 데이터를 존중하고 보호하며, 개인정보의 안전한 처리와 사용을 보장하여 프라이버시를 지켜야 한다.
 - 다) 안전과 보안: 인공지능 시스템의 안전한 운영 및 보안을 보장하여 해킹, 악용, 악의적인 공격으로부터 시스템을 보호해야 한다.
 - 라) 책임과 책임 소재: 인공지능의 사용, 개발, 배포, 결과에 대한 책임을 정의하고 적절한 책임 소재를 식별할 수 있어야 한다.
 - 마) 사회적 영향과 공익: 인공지능의 사용이 사회에 미치는 영향을 고려하고, 이를 통해 사회적인 공익을 증진시키는 방향으로 인공지능을 개발하고 활용해야 한다.
 - 바) 자율성과 의사결정 투명성: 자율적으로 의사결정하는 인공지능 시스템의 투명성을 유지하고, 의사결정 과정에 대한 이해 가능성을 보장해야 한다.
 - 사) 사람 중심적 가치: 인공지능 기술은 사람의 복리와 행복을 증진시키기 위해 사용되어야 하며, 사람 중심적인 원칙을 기반으로 개발되어야 한다.
 - 아) 지속 가능성: 인공지능 기술은 환경, 경제, 사회적 지속 가능성을 고려하여 개발하고 활용해야 한다.
- ✓ 이러한 윤리적 원칙은 인공지능 개발자, 이용자, 정부, 기업 및 기술 커뮤니티에게 인공지능의 윤리적인 사용과 발전을 촉진하며, 사회적으로 수용 가능한 방향으로 인공지능 기술을 유도해야 된다.

3.1.2.3. 인공지능 윤리 필요성

가) 필요성 인식

- (1) 인공지능 윤리를 논의해야 하는 부분은 인공지능을 활용에 따른 부작용에 대한 규범적 문제에 대해 법으로 대응하고자 하는 논의이다.

- (2) 즉, 인공지능에 대한 법적 규제 논의가 이미 활발하게 이루어지고 있으며 그동안 상당한 연구가 축적되어 있으나 윤리보다 더 강력하고 실효성 있는 제재 수단이라 할 수 있는 법적 규제로 인공지능 활용에 따른 부작용에 대한 규범적 문제를 충분히 다룰 수 있는 상태는 아니다.
- (3) 이는 법 개념에 비해 그 의미가 명확하지 않은 윤리 개념을 규제수단으로 활용할 수 있는 상태가 아니기 때문이다⁴⁷⁾.

나) 인공지능 윤리 필요성 대두

- (1) 인공지능기술의 급속한 발전으로 제조·의료·교통·환경·교육 등 산업 전 분야에 인공지능이 활용·확산되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 인공지능 윤리 이슈가 새롭게 대두되었다.
- (2) 인공지능 윤리에 관한 논란들이 다수 발생하고 있으며, 이를 기술오남용, 데이터 편향성, 알고리즘 차별, 프라이버시 침해 분야로 나누어 사례를 <표 3-1>과 같이 정리하였다.

<표 3-1> 인공지능 윤리 논란 사례

구분	시점	내용
기술 오남용	2019. 9	(보이스 피싱) 유럽 한 에너지기업의 CEO는 영국 범죄자들이 인공지능을 활용해 정교하게 만든 모회사 CEO의 가짜 음성에 속아 22만 유로를 송금하는 피해
데이터 편향성	2018. 10	(여성 차별) 아마존의 인공지능 기반 채용시스템이 개발자, 기술 직군 대부분에 남성만을 추천하는 문제가 발생함에 따라 아마존에서 동 시스템 사용 폐기
알고리즘 차별	2018. 1	(인종 차별) 인공지능 기반 범죄 예측 프로그램인 'COMPAS'의 재범률 예측에서 흑인 범죄자의 재범가능성을 백인보다 2배 이상 높게 예측하는 편향 발견
프라이버시 침해	2019. 9	정보 유출) 아마존 '알렉사', 구글 '구글 어시스턴트', 애플 '시리' 등이 인공지능 스피커로 수집된 음성 정보를 제3의 외부업체가 청취하는 것으로 밝혀져 논란

47) 인공지능 윤리와 인공지능 법학의 관계에 관해서는 김건우(2017), 로봇윤리 vs. 로봇법학: 따로 또 같이, 「법철학연구」, 제20권 제2호, 7-44쪽 참고.

- (3) OECD, EU 등 세계 각국과 주요 국제기구 등은 인공지능 윤리의 중요성을 인식하고 윤리적인 인공지능 실현을 위한 원칙들을 발표 중에 있다.
- (4) 우리나라에서도 2018년 4월「지능정보사회 윤리 가이드 라인(정보문화포럼)」, 2019년 11월「이용자중심의 지능정보사회를 위한 원칙(방통위)」 등이 발표되면서 본격적으로 논의되기 시작하였다.

다) 인공지능에 대한 법적 규제의 불충분성

- (1) 인공지능에 대한 법적 규제가 충분하지 않다는 점이다. 이는 인공지능이 야기하는 규범적 문제에 대응하는데 법적 규제는 완전하지 않다는 것으로 법적 규제가 갖고 있는 규범적 특징과 무관하지 않다.
- (2) 독일의 법철학자 클라우스 귄터(Klaus Günther)가 적절하게 지적한 것처럼 법, 특히 현대법의 바탕을 이루는 근대 법(modernes Recht)은 외부성이라는 특징을 지닌다⁴⁸⁾. 내면적 동기까지 문제 삼는 윤리나 도덕과는 달리, 법은 기본적으로 외부적에서 발생한 결과를 문제 삼는다는 것이다.
- (3) 칸트는 도덕과 법을 개념적·내용적으로 구분하면서, 내면적 동기와 자율성을 특징으로 하는 도덕과는 달리 법은 외부적 결과와 강제 가능성을 특징으로 한다고 말한다⁴⁹⁾. 이렇게 법은 기본적으로 수범자 또는 행위자가 외부적으로 야기한 결과를 문제 삼는다는 특징은 지금도 여전히 유지되고 있다. 물론 형법학의 논의가 보여주는 것처럼 행위자의 내면적 측면, 이를테면 고의 등도 이제는 행위가 불법인지 여부를 판단할 때 중요한 역할을 한다⁵⁰⁾. 그렇지만 법은 행위자에 의해 특정한 외부적 결과가 발생해야 비로소 이를 문제 삼는다는 특성은 여전히 유지된다.
- (4) 이러한 까닭에 법적 규제가 인공지능이 유발하는 규범적 문제를 완전하게 커버하기는 어렵다. 이를테면 시민들은 자율주행자동차가 특정한 결과를 일으키기 전에도 이러한 자율주행자동차에 적용된 알고리즘이 윤리적으로 볼 때 타당한지를 문제 삼을 수 있다. 바로 이러한 근거에서 인공지능에 대한 윤리 문제를 논의할 필요가 있는 것이다.

48) K. Günther, 김나경 (역)(2002), 형법의 대화윤리적 근거지움의 가능성 ; K. Gü이상돈 (역음), 대화이론과 법, 법문사, 152-154쪽 참고. 근대법에 관해서는 양천수(2009), 법의 근대성과 탈근대성: 하버마스와 투렌의 기획을 중심으로 하여, 『법학연구』, 제50권 제1호, 부산대학교, 161-191쪽 참고

49) I. Kant(1797), Metaphysik der Sitten, Einleitung in die Rechtslehre, §B, S. 337.

50) 이는 독일의 형법학자 벨첼(Hans Welzel)이 정립한 목적적 행위론에서 극명하게 드러난다. 이를 보여주는 H. Welzel, 황산덕(역)(1957), 형법체계의 신행상: 목적적 행위론 입문, 박영사.

라) 윤리를 통한 자율적 규제

- (1) 인공지능 윤리를 논의하는 이유는 인공지능이 유발하는 규범적 문제를 윤리라는 법보다는 강도가 약한 규제수단을 이용하여 규제하겠다는 의도에서 찾을 수 있다. 즉, 인공지능에 관한 규범적 문제에 윤리를 활용하는 자율적 규제로 대응하겠다는 것이다. 이는 법적 규제가 갖고 있는 강제성이라는 특징과 관련을 맺는다. 법적 규제는 외부 적으로 발생한 결과를 규제의 출발점으로 삼는 동시에 강제적인 집행방식을 갖추고 있는 직접적 규제방식이라는 특징을 지닌다.
- (2) 이로 인해 규범적 분쟁에 효과적으로 대응할 수 있다. 시간이라는 측면에서 보면, 법적 규제는 단기간에 목표로 하는 효과를 거둘 수 있다. 그렇지만 법적 규제는 규제의 강제성이라는 특징으로 인해 예기치 못한 부작용을 낳기도 한다. 이를 규제 이론가들은 규제의 역설이라고 표현한다⁵¹⁾.
- (3) 독일의 법사회학자 토이브너(Gunther Teubner)는 이를 조종의 트릴레마(regulatorisches Trilemma)라고 지칭하기도 하였다⁵²⁾. 그 때문에 법적 규제는 민법상 불법행위나 형법상 범죄와 같은 규범적 문제를 규제하기에는 적합하지만, 인공지능 개발과 같이 고도의 창의성을 요구하면서 과학기술 수준이 급격하게 발전하는 영역을 규제하기에는 적합하지 않다는 비판이 제기된다.
- (4) 이렇게 많은 것이 매순간 변하고 발전하는 영역에 대해서는 법적 규제를 철폐하거나 피규제자의 자율성을 전제로 하는 약한 규제(soft law)를 투입해야 한다는 것이다. 이러한 주장을 예증하는 것이 바로 요즘 화제가 되는 포괄적 네거티브 규제 전환이나 규제 샌드박스(regulatory sandbox)라 할 수 있다.
- (5) 인공지능 윤리 논의는 바로 이러한 맥락에서 이해할 수 있다. 인공지능처럼 현재 급속 하게 발전하고 있는 영역에 강력한 강제수단을 갖추고 있는 법적 규제를 광범위하게 투입 하는 것은 적절하지 않다는 것이다.
- (6) 따라서 한편으로는 인공지능 과학기술이 원활하고 창의적으로 발전할 수 있도록 하면서도, 다른 한편으로 인공지능이 유발하는 규범적 문제에 적절하게 대처하기 위해서는 법적 규제보다 직접성이나 강도가 약한 윤리적 규제를 고려할 필요가 있다는 것이다. 바로 이러한 맥락에서 인공지능과 윤리 문제를 논의할 필요가 있다.

51) 김영평, 최병선, 신도철 (편저)(2006), 규제의 역설, 삼성경제연구소.

52) Gunter. Teubner(1984), Verrechtlichung - Begriffe, Merkmale, Grenzen, Auswege, in: F. Kübler (Hrsg.), Verrechtlichung von Wirtschaft, Arbeit und sozialer Solidarität (Frankfurt/M.), S. 312 ff.

3.1.2.4. 인공지능 윤리에서 말하는 윤리 개념

가) 전제

- 인공지능 윤리에서 말하는 윤리 개념은 어떤 개념이어야 하는가?
- 공동체주의가 말하는 것처럼 도덕 개념까지 포괄하는 넓은 의미의 규범적 개념이 되어야 하는가?
- 자유주의가 말하는 것처럼 좁은 의미의 도덕 개념에 한정되어야 하는가?

나) 인공지능 윤리 개념

- (1) 인공지능 윤리에서 말하는 윤리는 자유주의가 강조하는 도덕 개념으로 한정하여 이해하는 것이 적절하다고 판단된다. 그 이유는 현재 우리에게 현실적으로 다가온 인공지능은 이른바 약한 인공지능(weak AI)이다.
- (2) 즉, 약한 인공지능은 아직 인간과 동일한 수준의 정신능력을 갖추지는 못한 인공지능을 말한다. 이러한 약한 인공지능은 인간과 같은 수준의 자율성을 행사할 수 없을 뿐만 아니라, 무엇이 좋은 것인지를 판단할 수 있는 능력도 갖추고 있지 않다.
- (3) 또한 약한 인공지능은 수단적인 존재는 될 수 있지만, 독자적인 존재는 될 수 없다. 무엇이 자신에게 좋은 것인지, 무엇이 자신을 행복하게 하는 것인지를 지각하거나 판단할 수 없다. 그러므로 이러한 약한 인공지능에게 포괄적인 의미의 윤리를 요구하는 것은 필요하지도, 바람직하지도 않다. 오히려 약한 인공지능에게 요청되는 것은 인간에 대한 관계에서 옳음을 준수하는 것, 이를테면 인간의 권리나 이익을 침해하지 않는 것이라 할 수 있다.
- (4) 위와 같은 점을 고려하면, 현 시점에서 인공지능에게 요구해야 하는 윤리는 그 의미 내용이 포괄적이면서도 모호한 윤리보다는 외연이 좁으면서도 그 의미 내용이 구체적인 도덕으로 보아야 한다. 이러한 측면에서 인공지능 윤리라는 표현보다는 인공지능 도덕이라는 표현이 더욱 정확하다.
- (5) 본 연구보고서에서는 인공지능 윤리라는 개념이 이미 널리 통용되고 있다는 점을 고려하여 인공지능 윤리라는 표현을 계속 사용하며, 이때 윤리는 도덕 개념을 지칭한다.

3.1.2.5. 윤리의 구별 개념

인공지능 윤리와 구별해야 할 개념이 있다. 사실 공학 영역에서 보면, 인공지능 윤리가 완전히 새로운 것은 아니다. 왜냐하면 그 이전부터 인공지능 윤리와 유사

한 논의가 공학, 더 나아가 과학기술 영역에서 전개되었기 때문이다. 이러한 예로 공학윤리를 들 수 있으며, 이외에도 인공지능 윤리와 유사한 논의로서 로봇 윤리를 언급할 수 있다. 따라서 이러한 공학윤리나 로봇윤리가 인공지능 윤리와 어떤 점에서 차이점이 있는 살펴보면 다음과 같다.

가) 공학윤리

- (1) 넓은 의미에서 인공지능 윤리 역시 공학윤리에 포함된다. 공학윤리는 공학 영역에서 공학적 설계를 수행하는 공학자에게 적용되는 윤리에 해당하는데, 인공지능 윤리 역시 인공지능을 설계하고 적용하는 개발자에게 적용되는 윤리라고 말할 수 있기 때문이다⁵³⁾.
- (2) 즉, 인공지능 윤리는 공학윤리 가운데 인공지능을 대상으로 하여 전문적으로 특화된 영역이라고 할 수 있는 것이다.
- (3) 그러나, 공학윤리에서는 공학 설계자를 윤리를 준수해야 하는 의무주체로 설정한다. 이때 공학 설계자에 의해 설계되고 만들어지며 운용되는 대상은 독자적인 주체라기보다는 철저한 수단으로 규정된다. 따라서 이러한 수단이 독자적으로 준수해야 하는 윤리는 생각하기 어렵다
- (4) 이와 달리 현재 진행되는 인공지능 윤리는 인공지능을 설계하는 설계자에게 적용되는 윤리뿐만 아니라 인공지능 자체가 준수해야 하는 윤리, 다시 말해 인공지능 자신을 위한 윤리까지 논의대상에 포함된다⁵⁴⁾. 이러한 측면에서 공학윤리와 인공지능 윤리는 구별된다.

나) 로봇 윤리

- (1) 로봇 윤리와 인공지능 윤리를 비교하면, 실질적으로 크게 구분되지 않은 채 논의되는 편이다. 로봇 윤리와 인공지능 윤리가 개념적인 측면에서 엄밀하게 구별되지 않은 채 거의 같은 의미로 다루어지고 있다. 즉, 로봇과 인공지능이 개념적·내용적으로 엄밀하게 구별되지 않은 채 논의되는 것과 맥락을 같이 한다. 이 점에서 아예 인공지능 로봇이라는 개념을 사용하여 논의를 전개하기도 한다⁵⁵⁾.
- (2) 그러나 엄밀하게 보면, 인공지능과 로봇은 개념적·내용적으로 구별해야 한다. 이는 독일의 사회학자 루만(Niklas Luhmann)이 정립한 체계이론의 관점을 수용할 때 분명해진다⁵⁶⁾.

53) 웬델 월러치·콜린 알렌, 노태복(옮김)(2016), 왜 로봇의 도덕인가, 메디치미디어.

54) 한국인공지능법학회(2019), 인공지능과 법, 박영사, p. 57.

55) 조성은 외(2018), 「인공지능시대 법제 대응과 사회적 수용성」, 정보통신정책연구원.

56) 체계이론의 관점을 수용하여 인공지능과 로봇을 구별하는 경우로는 양천수(2017), 인공지능과 법체계의 변

- (3) 이를테면 인공지능은 인간의 의식, 즉 의식체계를 인공적으로 구현한 것에 해당한다. 이에 반해 로봇은 인간의 몸, 즉 생명체계를 인공적으로 구현한 것이다.
- (4) 따라서 인공지능과 로봇은 서로 중첩되기도 하지만 그렇지 않은 경우도 많다. 예를 들어, 철저하게 인간에 의해 조종되는 로봇을 생각할 수 있다. 또는 알파고처럼 몸은 없고 지능만을 갖추고 있는 인공지능도 생각할 수 있다
- (5) 이러한 점을 고려하면, 인공지능과 로봇은 분명 구별된다. 그렇지만 오늘날 개발되는 로봇은 대부분 고도의 인공지능을 갖추고 있다. 그 때문에 인공지능하면 인공지능 로봇을 떠올리는 경우가 많다.
- (6) 이러한 이유에서 인공지능 윤리는 로봇 윤리와 중첩되어 논의된다고 말할 수 있다. 앞에서 논증한 것처럼, 인공지능과 로봇은 개념적·내용적으로 구별된다는 점에서 인공지능 윤리와 로봇 윤리 역시 구별된다.
- (7) 그렇지만 인공지능과 로봇이 결합되어 등장하는 오늘날의 상황을 고려할 때 양자를 거의 같은 의미로 사용하는 것도 큰 무리는 없어 보인다. 따라서 이 연구 보고서에서는 인공지능 윤리를 로봇윤리까지 포함하는 개념으로, 다시 말해 인공지능 로봇의 윤리라는 의미로 사용하고자 한다.

3.1.3. 인공지능 윤리의 수범 주체 및 기본 방향

3.1.3.1. 인공지능 윤리 관련자의 세 가지 유형

가) 인공지능 윤리 준수 주체

인공지능 윤리와 관련하여 정확하게 구별하고 판단해야 할 문제가 있다. 인공지능 윤리를 이행하고 준수해야 하는 주체, 즉 수범주체는 정확하게 누구인가 하는 문제가 그것이다.

이를 위해서는 먼저 인공지능 윤리와 관련을 맺는 주체, 다시 말해 인공지능 윤리를 준수해야 하는 주체에는 누가 있는지 검토해야 한다. 이에 관해서는 세 가지 유형의 주체를 거론할 수 있다.

(1) 첫째는 인공지능을 개발하는 개발자이다.

- ① 인공지능을 개발하는 개발자가 준수해야 하는 윤리로 이를테면 인공지능을 개발할 때 개발자는 인공지능이 사람의 권리나 이익을 침해하지 않도록 하거나 사람의 행복을 증진시킬 수 있도록 프로그래밍을 하고 알고리즘을 설계 및 적용해야 한다.

화: 형사사법을 예로 하여, 「법철학연구」, 제20권 제2호, pp. 45-76.

(2) 둘째는 인공지능 자체이다.

- ① 인공지능 자신이 준수해야 하는 윤리로 가령 인공지능 로봇은 로봇 3원칙에 따라 인간의 이익을 최우선적으로 고려하면서 행동해야 한다.
- ② 다만 이처럼 인공지능이 독자적인 윤리의 수범주체가 되기 위해서는 인간과 동등한 규범적 자율성을 인공지능이 보유하고 있어야 한다.
- ③ 다시 말하면, 강한 인공지능(strong AI)이 출현하는 것을 전제로 해야 한다.

(3) 셋째는 인공지능을 이용하거나 인공지능과 거래를 하는 이용자이다.

- ① 인공지능을 이용하거나 거래를 맺는 이용자가 준수해야 하는 윤리이다.
- ② 가령 인공지능을 이용하는 사람은 인공지능을 괴롭히거나 학대해서는 안 된다는 것 등을 들 수 있다.
- ③ 또한 인공지능을 함부로 폐기해서는 안 된다는 윤리도 생각해 볼 수 있다.

나) 인공지능 윤리 설정을 위한 핵심 주체

- (1) 현 시점에서는 인공지능을 개발하는 개발자를 우선적인 수범주체로 설정해야 한다고 판단되며, 이는 현재의 기술 수준으로는 아직 강한 인공지능을 개발할 수 없다는 가정이다.
- (2) 강한 인공지능이 언제 실현될 수 있을지도 예측하기 어렵기 때문에 인공지능 자체를 윤리의 수범 주체로 설정하는 것은 현재로서는 적절하지 않다.
- (3) 향후 인공지능 자체(강한 인공지능)를 법적 인격체로 인정할 수 있지만 법적 인격과 윤리적 인격은 개념적·내용적으로 구별된다. 또한 특정한 존재가 윤리적 인격체로 승인되기 위해서는 윤리적 자율성이라는 법적 자율성보다는 더욱 강력한 수준의 자율성이 요구되며, 이러한 부분은 고려하면 강한 인공지능이 가능하지 않은 현 상황에서는 인공지능을 독자적인 윤리적 수범주체로 설정하기는 어렵다.
- (4) 마지막으로 인공지능 이용자를 인공지능 윤리의 수범주체로 볼 수 있지만, 현 시점에 시급한 것은 아니다. 지금 상황에서는 인공지능 이용자에게 윤리를 부과하기보다는 인공지능으로부터 이들을 보호하는 방법을 모색하는 것이 더욱 시급하다.
 - ✓ 이와 같은 상황을 고려할 때, 인공지능 윤리 제정시 수범 주체는 현시점에서 인공지능 개발자 위주의 제정이 우선시 되어야 하며, 향후 인공지능 발전에 따라서 인공지능 자체 및 이용자에 대한 윤리 기준 및 법적인 인정 부분도 고려해야 한다.
 - ✓ 결과적으로 인공지능 윤리는 정확하게 말하면 인공지능 개발자가 준수해야 하는 윤리, 인공지능 개발자의 윤리라고 보는 것이 타당하다.

3.1.3.2. 인공지능 윤리 기본 방향

가) 공리주의·자유주의·공동체주의 윤리

철학이론의 차원에서 볼 때 인공지능 윤리는 공리주의 윤리, 자유주의 윤리, 공동체주의 윤리로 유형화할 수 있다⁵⁷⁾.

(1) 공리주의 윤리

(가) 공리주의의 관점에서 인공지능 윤리에 적용하여 도출하면 다음과 같으며, 여기서 공리주의는 기본적으로 양적 공리주의를 전제로 한다.

(나) 벤담이 주장한 공리주의는 공리를 극대화하는 것을 정의로 파악한다. 이는 최대다수의 최대행복으로 잘 알려져 있으며, 이러한 공리주의에서는 서로 다른 공리 간의 충돌과 형량, 달리 말해 이익형량을 정의에 도달하기 위한 중요한 방법으로 사용한다.

(다) 이익형량을 통한 공리의 극대화, 이것이 바로 공리주의가 취하는 기본 방법론이라 할 수 있다. 이러한 공리주의에 기반을 둔 인공지능 윤리는 주로 자율주행자동차 영역에서 발견되고 동시에 문제가 된다⁵⁸⁾. 이를테면 서로 다른 공리가 충돌되는 상황에서 자율주행자동차가 어떤 공리를 우선해야 하는지에 관해 논의가 전개된다.

- ① 자율주행 자동차 탑승자의 이익을 우선해야 하는지, 그게 아니면 보행자나 다른 자율주행자동차 탑승자의 이익을 우선해야 하는지가 문제된다. 이를 위해 윤리학에서 사용되는 유명한 기차 딜레마 사례가 자율주행자동차의 사례로 전환되어 논의된다⁵⁹⁾.
- ② 이때 이러한 딜레마적 상황을 해결하기 위해 자율주행 자동차를 설계하는 개발자는 어떤 윤리적 알고리즘을 적용해야 하는지가 논의의 초점이 된다.
- ③ 이에 관해 자율주행 자동차의 개발자나 회사 등은 자신들의 고객에 해당하는 탑승자의 이익을 우선하는 윤리적 알고리즘을 개발 및 적용하는 것으로 보인다.

57) 양해림 외(2018), 공학윤리, 충남대학교 출판부 양해림, p. 55.

58) 이종기·오병두(2016), 자율주행자동차와 로봇윤리: 그 법적 시사점, 『홍익법학』, 제17권 제2호, pp. 1-25.
 이상돈·정채연(2017), 자율주행자동차의 윤리화의 과제와 전망, 『IT와 법연구』, 제15집, pp. 281~325.
 변순용(2017), 자율주행자동차의 윤리적 가이드 라인에 대한 시론, 『윤리연구』, 제112호, pp. 199-216.
 변순용 외(2018), 자율주행자동차에 대한 한국형 윤리 가이드 라인 연구, 『윤리연구』, 제123호, pp. 203- 239.
 정승원·황기연(2018), 자율주행 자동차의 윤리적 선택에 따른 교통사고비용 분석, 『한국ITS학회 논문지』, 제17권 제6호, pp. 224-239.

59) 정진규(2016), 트롤리 문제와 다원론적 규범 윤리 이론, 『동서철학연구』, 제81호, pp. 423-446
 이주석(2017), 도덕직관에 관한 윤리적 해명 가능성: 트롤리 딜레마의 경우, 『철학논총』, 제90집, pp. 363~380.

- (라) 이와 같이 공리주의적·이익형량적 사고를 명시적으로 인공지능 윤리에 적용하는 시도는 일반 대중의 지지를 받지 못하고 있으며, 이는 일반적으로 우리 인간 존재는 자신이 이익형량의 대상이 된다는 사실에 심한 거부감을 갖고 있기 때문이다.
- (마) 따라서 이러한 공리주의적 윤리를 (명시적으로) 인공지능 윤리로 원용하는 것은 여전히 타당하지 않다.

(2) 자유주의 윤리

- (가) 자유주의는 행위주체의 자유를 가장 우선시 하는 윤리를 말하며, 자유주의 윤리는 정확하게 말하면 자유주의 도덕이라고 불러야 하며 이러한 자유주의 윤리는 두 가지로 구별할 수 있다.
- (나) 첫째 목적론적 자유주의: 목적론적 자유주의는 목적론, 달리 말해 공리주의의 관점에서 자유를 우선시 한다. 자유가 질적으로 볼 때 가장 우선시해야 하는 공리이기 때문이다. 이는 밀(John Stuart Mill)의 질적 공리주의에서 찾아볼 수 있다⁶⁰⁾.
- (다) 둘째는 의무론적 자유주의: 의무론적 자유주의는 목적론과는 무관하게 인간 주체의 자유를 우선시 하는 것이 우리 인간에게 부여된 의무라고 본다. 이는 칸트의 의무론적 도덕에서 발견할 수 있는데, 이후 대부분의 자유주의는 이러한 의무론적 자유주의를 수용한다⁶¹⁾. 자유를 목적이나 공리와 무관하게 가장 우선적인 규범적 가치로서, 이를 의무로 인정하는 것이다.
- (라) 이러한 의무론적 자유주의 또는 의무론의 견지에서 인공지능 윤리를 설정한 것으로서 SF 소설가 아이작 아시모프(Isaac Asimov)가 제시한 로봇 3원칙이 언급되며, 로봇 3원칙은 다음과 같은 세 가지 원칙으로 구성된다⁶²⁾.

- 제1원칙: 로봇은 인간에게 해를 입히거나 행동을 하지 않음으로써 인간이 해를 입도록 해서는 안 된다.
- 제2원칙: 로봇은 인간이 내리는 명령에 복종해야 한다. 다만 이러한 명령이 제1원칙에 위배될 때에는 예외로 한다.
- 제3원칙: 로봇은 자신의 존재를 보호해야 한다. 다만 이러한 보호가 제1원칙과 제2원칙에 위배될 때에는 예외로 한다.

- (마) 제1원칙은 인공지능 로봇은 작위 또는 부작위로 인간에게 해를 입혀서는 안 된다는 원칙이다.

60) 존 스튜어트 밀, 서병훈(옮김)(2018), 자유론, 책세상.

61) 임마누엘 칸트, 백종현(옮김)(2005), “윤리형이상학 정초”, 아카넷.

62) 로봇 3원칙의 번역은 웬델 윌러치·콜린 알렌, 노태복(옮김)(2016), 왜 로봇의 도덕인가, 메디치미디어, p. 13.

- ① 여기서 인간에게 해를 입힌다는 것은 인간의 권리 또는 인권을 침해해서는 안 된다는 의미로 새길 수 있다.
 - ② 즉, 인공지능 로봇은 인간의 권리를 침해해서는 안 된다는 것이다. 이는 밀이 제시한 해악의 원리(harm principle)를 반영한 것으로⁶³⁾ 이 점에서 로봇 3원칙이 자유주의를 반영하고 있음을 알 수 있다.
- (바) 제2원칙은 인공지능 로봇은 인간의 명령에 복종해야 한다는 것을 보여준다.
- ① 이러한 제2원칙에서 두 가지 의미를 읽어낼 수 있다. 먼저 인간과 인공지능 로봇의 관계가 어떤 관계인지 보여준다.
 - ② 제2원칙에 따르면, 인간과 인공지능 로봇의 관계는 수평적인 관계가 아니다. 이는 수직적인 관계이다.
 - ③ 구체적으로 말하면, 인공지능 로봇은 인간의 명령에 복종하는 마치 노예와 같은 존재가 되어야 한다. 여기서 인공지능 로봇 자체의 존엄성은 인정될 여지가 없다.
 - ④ 다음으로 인공지능 로봇이 인간의 명령에 복종해야 한다는 것은, 인공지능 로봇이 단순히 인간의 권리를 침해하지 말아야 하는 것을 넘어 인간의 의무에 복종함으로써 인간의 이익을 증진하는 데 적극적으로 기여해야 한다는 점을 보여준다.
 - ⑤ 이러한 점에서 보면, 제2원칙은 로봇 3원칙이 반드시 자유주의적 윤리에만 연결되는 것은 아님을 시사한다. 인공지능 로봇이 인간의 이익을 실현하는 데 적극 봉사해야 한다는 점은 자유주의 윤리에 친화적이기보다는 오히려 공동체주의적 윤리에 더 가깝기 때문이다.
- (사) 제3원칙은 인공지능 로봇이 자신을 스스로 보호해야 한다는 점을 보여준다.
- ① 인공지능 로봇에게 자기보존권을 인정하는 것이다. 이러한 제3원칙은 역설적인 의미를 갖는다.
 - ② 먼저 제3원칙은 인공지능 로봇에게 인간처럼 자기보존권을 인정한다는 점에서 자유주의적 윤리의 성격을 갖는다.
 - ③ 그러나 동시에 다음과 같은 의미를 가질 수도 있다. 만약 자신의 존재를 보호해야 한다는 것의 의미에 자신의 존재를 소멸시킬 수 없다는 의미 역시 포함시킬 수 있다면, 이는 역설적으로 자유주의적 윤리를 제한하는 공동체주의적 또는 후견주의적 윤리의 성격도 갖는다⁶⁴⁾.

63) 오병선(2006), 밀의 자유론과 해악의 원리, 『수행인문학』, 제36집, pp. 31-64.

Gerhard Seher, 이현준·홍영기(역)(2018), 원칙에 의한 형벌규범의 정당화 그리고 법익개념, 『형사법연구』, 제30권 제1호, pp. 207-234.

64) 권지혜(2016), 형법정책의 토대로서 자유주의적 후견주의에 대한 비판적 고찰: 드워킨(Gerald Dworkin)과

- (아) 이와 같이 로봇 3원칙을 전적으로 자유주의 윤리로만 파악하는 것은 타당하지 않을 수 있다. 그렇지만 제2원칙과 제3원칙이 보여주는 것처럼, 로봇 3원칙은 우선순위 규칙을 수용함으로써 자유주의적 성격을 유지한다.
- (자) 제3원칙보다는 제2원칙이 그리고 제2원칙보다는 제1원칙이 우선한다고 선언하고 있는 것이다.
- (차) 자유주의 윤리를 명확하게 선언하는 제1원칙이 가장 우선적인 지위를 차지함으로써 로봇 3원칙의 자유주의적 성격은 유지된다.
- (카) 그러나 이렇게 자유주의 윤리를 대변하는 로봇 3원칙은 다음과 같은 한계를 지닌다.
- ① 자율주행 자동차의 경우가 잘 보여주는 것처럼, 만약 서로 다른 사람의 이익과 이익이 서로 충돌하는 경우에는 어떻게 해야 하는지를 로봇 3원칙은 말하고 있지 않다는 점이다.
 - ② 가령 어느 한 사람의 이익을 우선시 하면 다른 사람에게 해를 미치는 이윤배반적인 상황에서 인공지능 로봇이 어떻게 행위해야 하는지를 규정하고 있지 않다.
 - ③ 이러한 상황에서 로봇 3원칙은 모두 제대로 작동할 수 없다. 여기에서 바로 로봇 3원칙, 더 나아가 자유주의 윤리의 한계가 노출된다.
 - ④ 서로 다른 사람의 자유 또는 권리와 이익이 상호적으로 충돌할 때 이를 어떻게 해결해야 하는지를 판단하는 원칙이 존재하지 않는 이상 자유주의 윤리는 온전하게 작동할 수 없다.

(3) 공동체주의 윤리

- (가) 공동체주의 윤리는 공동체 구성원이 모두 준수해야 하는 미덕(virtue) 또는 인륜성(Sittlichkeit)을 중시하는 윤리를 말하며, 이러한 미덕의 배후에는 좋음(the good)이 놓여 있다고 말한다.
- (나) 이때 주의해야 할 점은 오늘날 주장되는 공동체주의 윤리는 개인의 권리와 공동체의 미덕이 충돌할 때 무조건 공동체의 미덕을 우선시해야 한다고 주장하지는 않는다는 점이다.
- (다) 다만 이들이 강조하는 것은 철학적으로 볼 때 옳음(the right)에 바탕을 둔 권리도 궁극적으로는 좋음과 무관하지 않다는 것이다. 바로 그 점에서 이론적으로 볼 때 모든 개인적 권리는 궁극적으로는 공동체가 추구하는 좋음으로 귀결되고, 따라서 goodness를 대변하는 미덕이 우선적인 지위를 차지한다고 본다.

파인버그(Joel Feinberg)의 논의를 중심으로, 고려대 법학박사 학위논문.

- (라) 그렇지만 이러한 공동체주의 윤리는 상당히 추상적이고 개별 공동체의 특성에 의존하는 점도 강해 인공지능 윤리로 적용하기에는 적절하지 않다.
- (마) 예를 들어, 공동체적 미덕에 해당하는 행복(well-being)이나 선량한 풍속 등과 같은 개념은 그 의미의 외연이 너무 넓어 인공지능 윤리로 사용하기에 적합하지 않다. 우리 인간들 역시 구체적인 상황에서 무엇이 행복한 일이고, 무엇이 선량한 풍속에 합치하는지를 판단하는 게 쉽지 않기 때문이다.
- (바) 가령 동성혼을 인정하는 것이 공동체의 미덕에 합치하지 않는 것인지 판단하기 쉽지 않고, 이에 관해서는 격렬한 논쟁이 거듭되고 있다.
- (사) 아직은 약한 인공지능만이 기술적으로 가능한 현시점에서는 인간처럼 아주 복잡한 사고과정을 거쳐야만 판단할 수 있는 공동체주의 윤리를 인공지능 윤리로서 원용하는 것은 적절하지 않다고 말할 수 있다.

나) 기본 방향

(1) 자유주의 윤리

- (가) 먼저 인공지능 윤리는 자유주의 윤리를 지향해야 한다. 가장 우선적인 이유는 자유주의 윤리의 규범적 의미가 가장 구체적이고 명확하기 때문이다.
- (나) 자유주의 윤리는 사람과 사람의 관계를 문제 삼는다. 더욱 정확하게 말하면, 로봇 3원칙 가운데 제1원칙이 보여주는 것처럼 다른 사람의 자유나 권리를 침해하는 것을 금지한다.
- (다) 옳음과 좋음을 구별함으로써 윤리, 더욱 정확하게 말해 도덕이 무엇을 하고 무엇을 하지 말아야 하는지를 명확하게 명령한다. 그 점에서 자유주의 윤리는 아직 인간과 동등한 수준의 규범적 자율성을 갖추지 못한 약한 인공지능에게 프로그래밍하기 적합하다.
- (라) 앞에서 살펴본 것처럼, 자유주의는 목적론적 자유주의와 의무론적 자유주의로 구별할 수 있다. 그러나 이러한 구별은 인공지능 윤리에서는 큰 의미는 없다고 생각한다.
- (마) 목적론적 자유주의인가, 아니면 의무론적 자유주의인가 하는 문제는 각자가 존엄하면서 서로 평등한 지위에 있는 인간 존재 사이에서 의미가 있기 때문이다.
- (바) 그러므로 인간과 평등한 존재가 아니라 인간에 대해 도구적 존재에 해당하는 인공지능에게는 의무론과 목적론 간의 대립이 큰 의미를 갖지 않는다. 오히려 정확하게 보면, 인공지능에게 설정하고자 하는 자유주의 윤리는 목적론적인 성격을 갖는다고 말할 수 있다.

- (사) 왜냐하면 우리가 인공지능 윤리를 논의하는 이유는 이것이 인간에게 이익이 되기 때문이다. 그 점에서 인공지능에게 설정하고자 하는 자유주의 윤리는 목적론적 자유주의라 할 수 있다.

(2) 공리주의 윤리

- (가) 인공지능 윤리를 설정할 때는 이렇게 자유주의 윤리를 가장 우선시 하면서도 보완적으로 공리주의 윤리를 수용할 필요가 있으며, 공리주의 윤리를 배제한 채 자유주의 윤리만으로 인공지능 윤리를 설정할 수는 없기 때문이다.
- (나) 예를 들어, 사람과 사람 사이의 자유나 권리가 충돌하는 상황에서 서로의 이익을 최적화하기 위해서는 이익형량적 사고를 어느 정도 수용할 수밖에 없다. 이러한 문제 및 상황은 헌법학의 기본권 제한 문제에서 잘 드러난다⁶⁵⁾.
- (다) 물론 우리들이 공리주의에 갖고 있는 거부감으로 인해 인공지능 윤리에서 이익형량 원칙을 정면에서 수용하는 것은 어려울 것이다.
- (라) 이러한 문제를 해결하기 위해서는 이익형량 원칙을 규범적으로 발전시킨 비례성 원칙을 권리 충돌에 대한 해결 원칙으로 받아들이는 것을 고려할 수 있다.

(3) 공적 영역의 윤리와 사적 영역의 윤리 구별

- (가) 공적 영역의 윤리와 사적 영역의 윤리를 구별할 필요가 있으며, 인공지능은 공적 영역과 사적 영역 모두에서 사용될 뿐만 아니라, 장치 사적 영역에서 더욱 중요하게 사용될 것이기 때문이다.
- (나) 예를 들어, 공적 영역의 윤리로는 타인의 권리를 침해하지 말 것을 제1원칙으로 설정할 수 있는 반면, 사적 영역의 윤리로는 이용자의 명령에 복종할 것을 제1원칙으로 설정할 수도 있기 때문이다.
- (다) 이는 공법 영역에서는 민주주의와 법치주의가 강조되는 반면, 사법 영역에서는 사적 자치가 가장 중요한 원리로 자리매김 하는 것과 유사하다. 물론 공적 영역의 윤리와 사적 영역의 윤리를 어떻게 차별화할 것인지는 더욱 논의를 해야 할 필요가 있다.
- (라) 그렇지만 양자를 통일적으로 규율하고자 하는 인공지능 윤리를 목표로 하는 것은 적절하지 않다고 말할 수 있다.

65) 로베르트 알렉시, 이준일(옮김)(2007), 기본권이론, 한길사; 김도균(2008), 권리의 문법, 박영사.

3.1.3.3. 기본 방향 설정에 따른 윤리원칙

가) 인간의 존엄성 존중

- (1) 인간의 존엄성(Menschen- würde)이 구체적으로 무엇을 뜻하는지 문제될 수 있지만, 인간의 존엄은 가령 행복이나 선량한 풍속, 인륜성보다는 그 개념이 명확한 편이다⁶⁶⁾.
- (2) 즉, 인간의 존엄에 관해서는 칸트가 정언명령으로써 비교적 정확한 공식을 제시하였으며, 이에 따르면, 인간을 목적이 아닌 철저한 수단으로 이용하는 경우가 인간의 존엄성을 침해하는 것이 된다.
- (3) 그러므로 만약 인공지능이 이용자가 아닌 다른 사람을 철저하게 수단으로 취급한다면 이는 그 사람의 존엄성을 침해하는 것이 될 수 있다.
- (4) 다만 이 경우에도 과연 어떤 경우가 인간을 철저하게 수단으로 이용하는 경우인가가 정의가 필요하므로 인간의 존엄성이 침해되는 경우를 유형화하여 이를 프로그래밍할 필요가 있으며, 인간의 존엄성 원칙은 가급적 보충적으로 적용할 필요가 있다.

나) 권리 침해 금지

- (1) 인공지능은 인간의 권리를 침해하지 말아야 한다. 이때 말하는 권리는 기본적으로 법적 권리를 말하지만, 단순히 법적 권리에만 그 의미를 한정해서는 안 된다.
- (2) 만약 그렇게 하면, 권리침해 금지 원칙은 윤리적 원칙이 아닌 법적 원칙이 될 것이다. 따라서 인공지능이 존중해야 하는 인간의 권리 개념에는 법적 권리뿐만 아니라 도덕적 권리, 가령 아직 실정화되지 않은 인권까지 포함시켜야 한다.
- (3) 이렇게 보면, 권리를 침해하지 말아야 한다는 원칙은(실정화된 인권과 실정화되지 않은 인권을 모두 포함하는 의미의) 인권을 침해하지 말아야 한다는 원칙으로 새길 수 있다.

다) 실정법 위반 금지

- (1) 인공지능은 실정법을 위반하지 말아야 한다. 보통 윤리는 실정법보다 그 외연이 넓은 규범이므로, 윤리를 준수한다는 것은 바로 실정법을 준수한다는 의미를 포함한다.
- (2) 특별한 예외를 제외하면, 실정법을 위반하는 행위가 윤리를 준수하는 것이라고 말할 수는 없다. 그러므로 인공지능은 실정법이 규정하는 각종 원리나 권리, 의

66) 베르너 마이호퍼, 심재우·윤재왕(윌김)(2019), 법치국가와 인간의 존엄, 세창출판사.

무 등을 준수해야 한다.

- (3) 이를테면 자율주행 자동차는 당연히 도로교통법을 준수하면서 운행해야 한다. 노동현장에서 사용되는 인공지능은 각종 노동관계법을 준수해야 한다. 의료현장에서 사용되는 인공지능은 의료법을 준수하면서 치료행위를 해야 한다.

라) 이용자의 명령 복종

- (1) 인공지능은 자신을 이용하는 사람이 내리는 명령에 복종해야 한다. 즉, 실정법 금지 원칙이 공적 영역에서 인공지능이 준수해야 하는 윤리라면, 이용자의 명령에 복종해야 한다는 원칙은 사적 영역에서 인공지능이 준수해야 하는 윤리에 해당한다.
- (2) 이는 커스터 마이징된 인공지능 윤리라 할 수 있다. 다만 인공지능이 이용자가 내리는 모든 명령에 복종해야 하는 것은 아니다.
- (3) 예컨대, 이용자가 내린 명령이 실정법이나 다른 윤리원칙을 위반하는 경우에는 이에 거부할 수 있도록 해야 한다. 인공지능이 다른 사람의 권리를 침해하라는 명령을 준수해서는 안 될 것이다.
- (4) 이 점에서 이용자의 명령 복종 원칙에 대한 한계로서 인공지능에게 명령 복종 거부권을 인정해야 한다.

마) 투명성

- (1) 인공지능은 자신이 내린 일체의 판단 및 결정에 관해 투명성을 유지해야 한다. 이때 투명성(transparency)이란 두 가지 의미로 구체화할 수 있다.
- (2) 우선 인공지능이 특정한 판단이나 결정을 내릴 때는 공식적으로 미리 주어진 요건과 절차를 준수해야 한다는 것이다. 즉, 인공지능이 내린 판단이나 결정이 절차적으로 정당화될 수 있어야 한다는 것이며, 더 나아가 인공지능이 과연 어떤 근거에서 판단이나 결정을 내린 것인지 설명할 수 있어야 한다는 것이다.
- (3) 여기서 두 번째 의미인 설명가능성은 그 자체 중대한 의미를 지니기에 이는 독자적인 윤리원칙으로 설정할 필요가 있다.

바) 설명 가능성

- (1) 인공지능은 자신이 내린 판단이나 결정이 어떤 절차 및 근거에 의한 것인지를 설명할 수 있어야 한다.
- (2) 이는 주로 인공지능이 특정한 판단이나 결정을 내릴 때 사용된 알고리즘이 어떻게 구성되어 있는지를 설명함으로써 실현된다. 이러한 설명가능성 원칙은 위

에서 언급한 투명성 원칙에 포함된 규범적 의미이다. 동시에 설명가능성 원칙은 투명성 원칙과 더불어 절차적 윤리에 해당한다.

- (3) 이러한 설명가능성 원칙은 전체 인공지능 윤리원칙 중에서도 결코 적지 않은 비중을 차지할 뿐만 아니라, 인공지능이 스스로 인공지능 윤리를 준수할 수 있도록 하는 일종의 윤리이행 수단이 된다는 점에서 독자적인 윤리원칙으로 설정할 필요가 있다.

사) 책임

- (1) 인공지능은 자신이 내린 판단이나 결정이 낳은 결과에 책임을 져야 한다. 이러한 책임 원칙은 사후적 윤리에 속한다. 물론 책임 원칙이 법이 아닌 윤리 영역에서 실제적인 의미를 가질 수 있을지 의문을 제기할 수 있다.
- (2) 그렇지만 책임 원칙은 윤리 영역에서도 여전히 실천적인 의미를 갖는다. 먼저 자신이 내린 판단이나 결정으로 인해 윤리적으로 좋지 않은 결과가 발생하였을 때 인공지능은 이에 사죄를 표명할 수 있다. 나아가 자신이 야기한 결과가 특정한 손해를 야기하였을 때는 법이 명령하기 이전에 자발적으로 손해를 배상할 수도 있다.
- (3) 물론 엄밀하게 말하면, 이때 손해배상을 하는 주체는 인공지능 자신이 되기보다는 인공지능 개발자나 소유자 등이 될 것이다. 인공지능이 자신의 명의로 책임재산을 갖고 있다고는 생각할 수 없기 때문이다.

3.1.4. 대한민국 인공지능 윤리 기준

2020년 12월 23일 과학기술정보통신부는 대통령 직속 4차산업혁명위원회 전체 회의에서 사람이 중심이 되는 「인공지능 윤리 기준」을 심의·의결했다. 이는 해외 각국에서 다양한 인공지능 윤리원칙이 발표되는 추세에 발맞춰 우리나라의 인공지능 국가전략에 따라 국가차원의 인공지능 윤리 기준을 마련하기 위함이다.

특히 윤리적 인공지능을 실현하기 위해 정부·공공기관, 기업, 이용자 등 모든 사회구성원이 인공지능 개발~활용 전 단계에서 함께 지켜야 할 주요 원칙과 핵심 요건을 제시하는 기준으로써 인공지능·윤리학·법학 등 학계·기업·시민단체를 아우르는 주요 전문가들이 자문과 의견수렴 과정을 거치며 윤리철학의 이론적 논의와 연계하여 ‘인간성을 위한 인공지능(AI for Humanity)’을 목표로 하는 윤리 기준을 마련한 것이다.

이를 통해 인공지능 윤리 기준은 모든 사회구성원이 인공지능개발 및 활용 전

단계에서 함께 지켜야 할 주요 원칙과 핵심 요건으로 ‘인간성(Humanity)’을 최고 가치로 내세운다. 그에 맞춰 인간 존엄성 원칙·사회의 공공선 원칙·기술의 합목적성 원칙으로 구성된 3대 기본원칙과 인권 보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안정성, 투명성의 10대 핵심요건을 제시하였다.

‘사람이 중심이 되는 「인공지능 윤리 기준」’은 윤리가 기술 발전을 저해하지 않음과 동시에 민간의 자율적인 노력을 존중하기 위한 새로운 사회규범으로서 ‘사람 중심’이라는 인공지능 개발과 활용의 방향성을 제시하여 인공지능의 도입과 확장으로 커지는 우려를 완화하고 파생되는 역기능을 최소화하기 위해 제시하고자 했다는 특징이 있다. 특히 특정 분야에 제한받지 않고 사회 구성원 모두가 활용할 수 있는 보편적인 윤리원칙으로 구성되어 있다.

특히 인공지능 생명주기에 걸쳐 반드시 충족되어야 하는 핵심요건을 도출하여 설계부터 활용 단계까지 인공지능에 관한 윤리적인 기준을 제시하고 있다⁶⁷⁾.

3.1.3.1. 3대 기본원칙 : 인공지능 개발 및 활용 과정에서 고려될 원칙

‘인간성을 위한 인공지능(AI for Humanity)’을 위해 인공지능 개발에서 활용에 이르는 전 과정에서 고려해야 할 3대 기본원칙을 제시

가) 인간 존엄성 원칙

- (1) 인간은 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계제품과는 교환 불가능한 가치가 있다.
- (2) 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용되어야 한다.
- (3) 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.

나) 사회의 공공선 원칙

- (1) 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- (2) 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- (3) 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관

67) 문정욱 외(2022. 12), AI 윤리체계 마련, “인공지능 윤리체계 확립을 위한 정책연구”, 과학기술정보통신부.

점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

다) 기술의 합목적성 원칙

- (1) 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.
- (2) 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.

3.1.3.2. 10대 핵심 요건 : 기본 원칙을 실현할 수 있는 세부 요건

3대 기본 원칙을 실천하고 이행할 수 있도록 인공지능 전체 생명주기에 걸쳐 충족되어야 하는 10가지의 핵심 요건을 제시하였다.

가) 인권보장

- (1) 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 권리를 보장하여야 한다.
- (2) 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

나) 프라이버시 보호

- (1) 인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.
- (2) 인공지능 전 생애주기에 걸쳐 개인정보의 오용을 최소화하도록 노력해야 한다.

다) 다양성 존중

- (1) 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영해야 하며, 성별·연령·장애·지역·인종·종교·국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다.
- (2) 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

라) 침해금지

- (1) 인공지능을 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안 된다.
- (2) 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

마) 공공성

- (1) 인공지능은 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- (2) 인공지능은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다.
- (3) 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.

바) 연대성

- (1) 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다.
- (2) 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다.
- (3) 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

사) 데이터 관리

- (1) 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- (2) 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다.

아) 책임성

- (1) 인공지능 개발 및 활용과정에서 책임주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- (2) 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다.

자) 안전성

- (1) 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- (2) 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

차) 투명성

- (1) 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상

- 황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- (2) 인공지능 기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

3.2. 인공지능 부작용 사례(윤리적 측면)

3.2.1. 인공지능 관련 이슈

3.2.1.1. 인공지능 투명성에 따른 이슈

- 가) 인공지능 기반 채용 솔루션을 제공하는 HireVue사의 제품은 인공지능 면접 과정에서 사용되는 안면인식 기술과 알고리즘이 검증되지 않았다는 이유로 투명성, 공정성, 그리고 책무성과 관련하여 탄원서가 연방거래위원회(FTC)에 제출하였다(Forbes, '20. 2).
- 나) 이후, HireVue사는 해당 소프트웨어의 주요 기능인 인터뷰 녹화를 바탕으로 면접자의 표정을 분석하는 기능을 삭제하였다.
- (1) 예 1) 인공지능 알고리즘의 블랙박스 특성으로 인한 책임소재 규명의 어려움: 인공지능의 의사결정에서 오류발생 시 딥러닝의 특성상 오류발생 여부 및 발생 지점 등을 판단하기 어려워 피해가 발생하여도 책임 소재 규명이 어려움.
- (2) 예 2) 금융 분야에서 활용되는 로보어드바이저는 인공지능 알고리즘을 활용하여 고객의 자산을 운용 및 관리해주는 자동화 서비스로 알고리즘의 오류·결함으로 인한 이용자 피해 발생 시 책임소재 규명 및 감시체계가 부재한 상황임. 인공지능 알고리즘이 블랙박스임을 이유로 내세워 책임회피 가능성이 있으므로 피해 발생 시 이의제기를 할 수 있는 담당 부서나 책임자가 그리고 이에 대한 법제가 필요함.

3.2.1.2. 인공지능 공정성에 따른 이슈

- 가) 영국 대입 시험 점수산정 시스템: 2020년 영국 시험감독청은 코로나19로 취소된 대입 시험 대신 알고리즘으로 성적을 산출했으나, 편향된 알고리즘의 결과에 대해 강한 사회적 반발이 일어나자 인해 이를 철회⁶⁸⁾ 하였다.
- 나) 한국 인공지능 챗봇 ‘이루다’ 이슈: 스타트업 ‘스캐터랩’의 챗봇 ‘이루다’는

68) <https://www.joongang.co.kr/article/23875682#home>

데이터 구축·학습 과정에서의 개인정보유출, 장애인·성소수자·인종 등에 대한 혐오 발언, 이용자의 챗봇 성희롱 대화로 젠더 편견 강화 등 사회적 논란을 빚어 2021년 서비스를 중단하였다.

- 다) 미국 경찰 범죄 수사용 안면인식 시스템 이슈: 미국 경찰국은 안면인식 기술을 활용한 범죄 수사 사건에 흑인 3명을 범죄자로 오인하여 부당 고소·체포함으로써 인권침해와 인종차별 논란이 있었으며, 미국 매사추세츠공대(MIT) 미디어랩 연구팀은 2018년 미국 뉴욕에서 열린 기계학습연구학회에서 인공지능에 활용되는 데이터가 백인과 남성 중심으로 구성되어 있어 피부색이 어두울수록 안면인식 인공지능의 오차율이 높다고 밝혔다.

3.2.1.3. 인공지능 오작동에 따른 이슈

- 가) 테슬라를 비롯한 구글, 아우디 등 자동차 제조 기업은 오토파일럿 기능을 갖춘 자율주행 자동차 개발 중이나 오토파일럿으로 인한 사고가 지속적으로 발생하여, 안전성을 확보하기 위한 법적·제도적·윤리적 정비 그리고 사용자의 행동 패턴을 고려한 설계가 필요하다.
- 나) 미국 애리조나주에서 시범운행 중이던 우버 자율주행차가 횡단 보도 밖에서 무단횡단 보행자를 치어 사망한 사례(2018.3)가 있었으며, 이는 인공지능의 발달 단계가 인간과 비인간 사물을 정확하고 세밀하게 구분하지 못함에 기인한다.
- 다) 한국의 한택투자증권에서 사용한 알고리즘 자동 매매 시스템은 직원의 이자율 입력 실수를 그대로 받아들여 거래를 자동 진행하여, 약 460억 원의 손실이 발생하여 그 결과로 회사가 파산⁶⁹⁾(2018.4)하게 되어, 예상 범위를 벗어난 거래에 경고창을 뜨게 하거나 주문을 자동으로 닫는 위험방지 장치가 도입되었다.

3.2.1.4. 인공지능 보안 위협에 따른 이슈

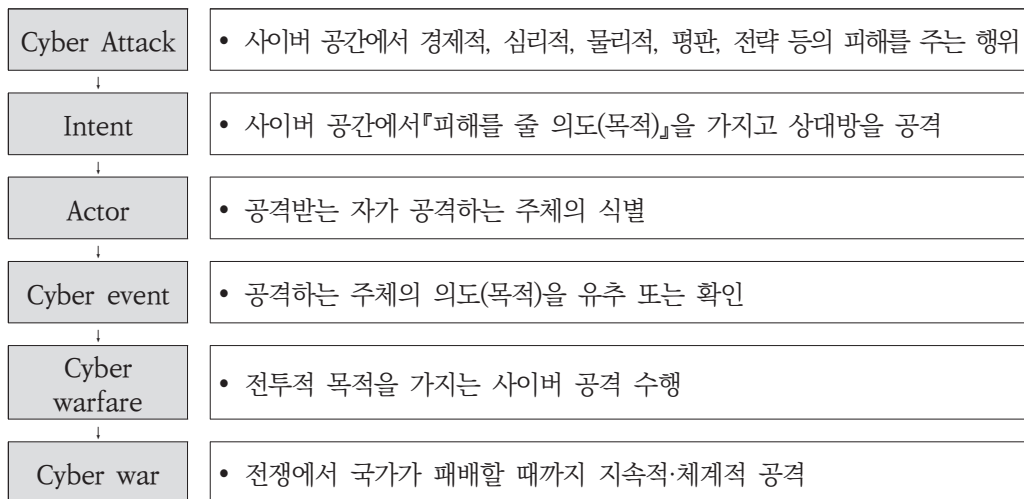
- 가) 인공지능시스템은 IoT(사물인터넷) 환경에서 컴퓨터와 통신망을 이용하므로 외부의 사이버 공격에 취약하다.
- ※ 대표적인 예로 실시간 도로 상황을 반영하는 자율주행 차량은 외부의 사이버 공격으로 원격 제어와 자율주행이 방해받을 수 있고, 이로 인한 위협으로 고의적 사고 야기와 국가 주요 인물의 납치 및 테러가 가능

69) https://biz.chosun.com/site/data/html_dir/2018/04/09/2018040900258.html

- 나) 따라서 자율주행 차량, 인공지능 교통 시스템, 실시간 데이터를 활용하는 군사 무기는 해킹으로 인한 오작동의 경우 심각한 상황으로 이어지므로 강력한 보안 기능이 함께 탑재되어야 한다.
- 다) 크라이슬러사는 차량 내 탑재된 시스템에 대한 공격으로 인해 자동차의 운행과 브레이킹을 제어할 수 있다는 취약점을 인식하여 자동차 140만 대 리콜(2015)

3.2.1.5. 사이버 공격 대응 부재 및 부족

- 가) 미군에서는 사이버전을“적의 교란, 마비, 파괴로부터 자기 자신의 정보체계의 통합성을 보존하는 동시에 적의 정보체계를 교란, 마비, 파괴하는 행위와 군사력을 운용하는 데 있어서 정보우위를 달성하는 절차”로 정의⁷⁰⁾하였다.
- 나) 정보화 시대의 전쟁개념은 산업기반시설, 국가 기간산업 및 군사체계에 대한 물리적 파괴(hard kill)보다는 기능적 마비(soft kill)에 초점이 두어질 것이며, 전차, 유도탄, 항공기, 함정 등이 수행하던 기존의 전쟁방식에 추가하여 컴퓨터의 가상공간만을 전장으로 하는 새로운 전쟁 영역을 형성⁷¹⁾되었다.
- 다) 사이버전은 6단계의 과정으로 발전⁷²⁾



- 라) 사이버전에 대응하기 위해서는 공격 시나리오를 정의하고, 이에 대응할 수 있는 체계를 마련하는 것이 필요⁷³⁾하다.

70) Park, H. K. (2013). Types and Information Security Technology on Cyber Warfare. Review of Korea Contents Association, 11(4), pp.41~44.

71) 노훈, & 이재욱. (2001). 사이버전의 출현과 영향, 그리고 대응방향. 국방정책연구, 가을호.

72) Robinson, M., Jones, K., & Janicke, H. (2015). Cyber warfare: Issues and challenges. Computers & security, 49, pp.70~94.

- 마) 이러한 전략적인 접근을 위해 빅데이터 기반 현황 정보 수집 및 인공지능 기반 공격 이벤트 연관 분석 기술 적용한 사이버전 대응 시스템을 구축⁷⁴⁾해야 한다.
- 바) 사이버전 대응을 위한 인공지능 기반 대응 플랫폼에 Overfit을 적용하면, 변조 공격을 차단하는 것이 불가하다.
- 사) 반대로 Underfit을 적용하면 모든 변조 공격 차단할 수 있지만, 정상적인 접근도 차단되는 문제점이 발생한다.

3.2.1.6. 전쟁 발발 가능성 증대

- 가) 자율무기체계 비용은 전투원 양성 비용 대비 그 비용이 매우 낮으며, 무기 수량 제한 부재로 양산이 용이하다.
- 나) 군인과 일반 시민, 정치인과 정책결정권자가 실제 전장 터의 참상을 직접 경험할 기회를 배제될 수 있다.
- 다) 상황에 따라, 추상적인 기대와 희망에 기초하여 전쟁 여부 결정이 가능하다.
- 라) 병력 사망 위험이 없어 심리적 부담 감소에 따른 자율무기체계 배치 및 활용 가능성 증대⁷⁵⁾될 수 있으며, 이에 따른 국가 간 무력 분쟁 발발 가능성이 증대될 수 있다.

3.2.1.7. 자율무기체계의 목적 외 용도에 사용

- 가) 자율무기체계를 개발하는 기관 중에는 민간 기업도 존재하며, 이러한 기업이 자율무기체계를 국제 범죄 카르텔이나 비국가적 행위자에게도 판매할 가능성이 있다.
- 나) 국제 범죄 카르텔이 자율무기체계를 취득 후 사용할 경우, 국가 주요 인물 암살이나 시설 파괴가 용이하게 된다.

73) 안예찬, 신영현, & 이근호. (2015). 빅데이터 플랫폼과 모니터링 시스템의 융합을 이용한 BLE 기반의 ZEP 시스템 공격 기법에 대한 대응방안 연구. *Journal of Digital Convergence*, 13(8), pp.331~336.

74) Kim, J. H., Lim, S., Kim, I. K., Cho, H. S., & No, B. K. (2013). Technical trends of cyber security with big data. *Electronics and Telecommunications Trends*, 28(3), pp.19~29.

75) Musandu, N. N. (2016). Doctor of Philosophy Thesis in Military Informatics (OpenPhD): Lethal Autonomy of Weapons is Designed and/or Recessive.



〈그림 3-2〉 터키 STM사, KARGU Drone, Dec. 2019.

3.2.1.8. 생명 경시 및 인간 존엄 훼손

가) 자율무기체계의 판단 방법 : 기계가 인간의 개입 없이 목표물을 선별하거나 (결정) 공격하는(행동) 것 자체가 아니라, 무기 설계자가 살상할 결정 (decision to kill)을 인간에서 기계로 기능적으로 부여했는지 여부⁷⁶⁾

✓ 기계가 적군 살상 여부를 결정할 때, 이러한 상태를 자율무기체계로 판단

나) 전투 불가능하거나 전투 의지가 없는 상태의 적군을 확인하였을 때, 기계가 살상 결정 가능

다) 전투 의지가 없는 상태의 적군의 삶에 대한 권리 무시 가능

라) 인간성이 배제된 기계에 자율적 판단에 의한 자동 살인이 인간의 존엄성 침해

✓ 이에 따라, UN은 “어떤 경우에도 감정을 갖지 않는 살인 로봇 병기에게 생명의 생사 결정 권한 배제 주장”

✓ 세계적으로 살인 로봇이 확산되기 전에 관련 연구개발의 일시 정지를 권고안으로 제시⁷⁷⁾

76) 김자희, 장신, & 주성구. (2017). 자율 로봇의 잠재적 무기화에 대한 소고-개념정립을 통한 규제를 중심으로. 입법과 정책, 9(3), pp. 135~156.

77) 한희원. (2018). 인공지능(AI) 치명적자율무기(LAWs)의 법적·윤리적 쟁점에 대한 기초연구. 중앙법학, 20(1), pp. 325~365.

3.2.1.9. 기타 인공지능의 부작용 사례

- 가) 인공지능은 인간의 편익을 증대시키지만, 그에 못지않게 위험도 증대시키고 있다. 2017년 Global Risk Report에 따르면 12개의 새로운 기술 중 인공지능과 로봇 공학이 편익과 위험 모두 가장 큰 것으로 조사⁷⁸⁾되었다.
- 나) 인공지능이 유발하는 위험 가능성에는 아래와 같이 편견, 차별, 사생활 침해 등이 있다.

인공지능이 유발하는 위험 가능성 ⁷⁹⁾	
• 편견과 차별 • 개인의 자율성·의지·권리 보장에 대한 부정 • 불투명하거나, 불명확하거나 혹은 정당하지 않은 결과 초래	• 사생활 침해 • 사회적 관계의 고립과 붕괴 • 신뢰할 수 없거나 위험하거나 혹은 질적으로 낮은 수준의 결과

- 다) 인공지능의 위험사례는 먼저, 아마존의 채용시스템에서 확인할 수 있다. 아마존은 2014년 인공지능을 이용한 채용시스템을 활용하였지만, 여성을 차별하는 알고리즘이 발견되어 2015년도에 해당 시스템을 폐기하였다.
- 라) 다음으로, 미국에서 널리 사용되는 재범 위험예측 알고리즘인 COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)는 흑인과 백인 간의 재범 예측률에 있어서 흑인에게 불리한 결과를 나타내었다.
- 마) 또한, 최근에 아마존에서 개발한 안면인식 기술이 미국 국회의원 28명을 범 죄자로 잘못 인식한 사례⁸⁰⁾도 있었다.

3.2.1. 군사용 인공지능 자율무기체계의 과대/과소 적합에 따른 문제점

3.2.1.1. 인공지능 자율무기체계의 과대/과소 적합

- 가) 인공지능은 생성된 학습 결과(Learning model)가 예상 모델(Expected model)과의 차이 정도에 따라 과대 적합(Overfit), 과소 적합(Underfit)으로 구분⁸¹⁾된다.

78) 오윤경 외 3인, '혁신과 위험관리: 사람중심 기술혁신을 위한 추진과제' ISSUE PAPER, 한국행정연구원, 2020. 6. 9. p. 9.

79) Yang Jongmo, 인공지능의 윤리/정책/이슈, ISSUE PAPER No. 08, KAIST, 2019, 12. p.6.

80) Jacob Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots", American Civil Liberties Union(ACLU), 2018. 7. 26.

81) Muslim, M. A., & Dasril, Y. (2021). Company bankruptcy prediction framework based on the most

Learning Model	Diagram	설명
과대 적합 (Overfit)		• 학습 모델을 생성할 때, 학습 모델 정확도에 너무 집중한 상태(High Variance, 과분산) • 예상 학습 결과(Expected model)와 실제 학습 결과(Learning model)의 과도한 차이가 있음 • 모델 적용에 따른 예측(Predict)할 때, '해당(Match)로 판단해야 하지만 학습 모델과의 차이로 '해당 없음(Not match)'으로 판단
과소 적합 (Underfit)		• 학습 모델을 생성할 때, 학습 모델 사용률에 너무 집중한 상태(High Bias, 과편향) • 실제 학습 결과(Learning model)가 예상 학습 결과(Expected model)를 상당한 여유를 가지고 포괄 • 모델 적용에 따른 예측(Predict)할 때, '해당 없음'으로 판단해야 하지만 '해당'으로 판단

3.2.1.2. 자율무기체계에 인공지능 적용의 어려움

가) 과대 적합·과소 적합은, 학습 반복 수준과 학습 데이터의 분포에 따라 그 결과가 다르게 나타난다⁸²⁾.

나) 인공지능을 적용하는 분야에 따라 요구되는 학습 결과가 달라질 수 있다⁸³⁾.

다) 이러한 과대/과소 적합 현상을 방지하기 위해서는 충분한 학습데이터와 이를 평가할 수 있는 데이터 분석 전문가, 그리고 이를 활용하기 위한 분야 (Domain) 전문가 필요⁸⁴⁾하다.

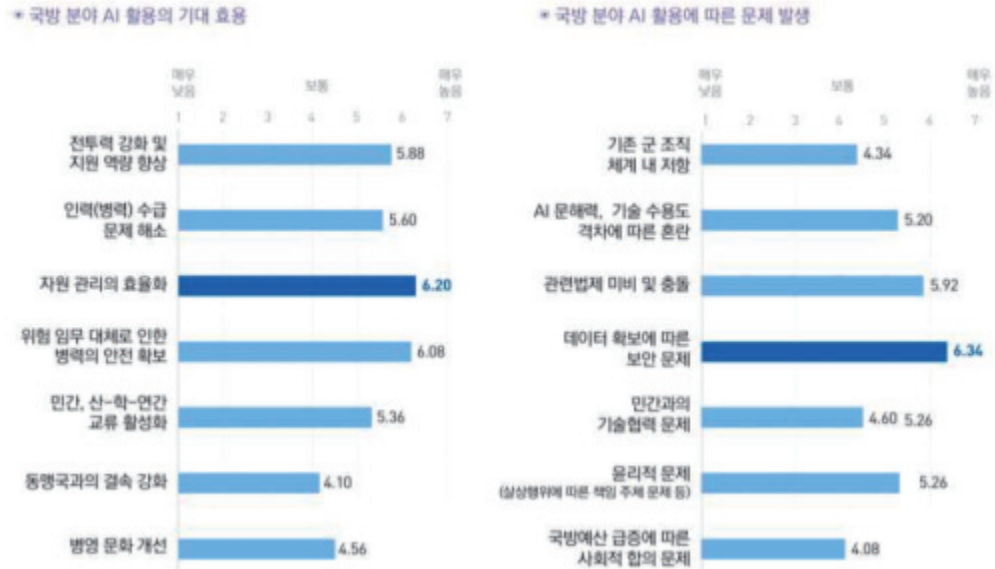
influential features using XGBoost and stacking ensemble learning. International Journal of Electrical & Computer Engineering (2088-8708), 11(6).

82) Czajkowski, M., & Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. Expert Systems with Applications, 137, pp.392~404.

83) Singh, J., Walden, I., Crowcroft, J., & Bacon, J. (2016). Responsibility & machine learning: Part of a process. Available at SSRN 2860048.

84) Yannakakis, G. N., & Togelius, J. (2018). Artificial intelligence and games(Vol. 2, pp.

라) 그러나 국방 분야의 특징상, 일부 분야(ex. 전장, 전투 방식 등)에 대한 학습 데이터를 확보하기 어렵기 때문에⁸⁵⁾, 인공지능 적용 분야를 고려할 필요가 있다.



〈그림 3-3〉 국방분야 AI활용에서의 기대 효용 부문 및 발생 가능한 난제

3.2.1.3. 인위적인 인공지능 과대 적합·과소 적합 적용

- 가) 일반적인 인공지능 분야에서는 과대/과소 적합을 조율하여 적정 학습 모델을 생성하는 것이 중요하다.
- 나) 특히 자율무기체계에서는 무기체계의 종류와 목적에 따라 강조되는 항목이 상이하므로 상황에 따라 학습 모델을 인위적으로 과대/과소 적합을 유도하여, 학습 모델의 활용 가능성을 높일 필요가 있다.

Learning Model	요구사항	사례
과대 적합 (Overfit)	- 발생 Event가 목표와 일치하는지를 확인 - 임무에 집중하여, 여타 부가적인 피해 최소화	- 폭발물 탑재 드론을 사용한 건물, 시설 파괴 - 이동 중인 인물 식별
과소 적합 (Underfit)	- 다소 시간과 비용이 들더라도 임무 달성에 집중 - 안전한 환경을 구축하거나 무기나 장비의 가용성을 유지하는데 집중	지뢰 제거 로봇에 의한 특정 구역에 설치된 지뢰 제거

2475-1502). New York: Springer.

85) 윤정현. (2020). 국방 분야의 인공지능 활용성 제고 방안과 시사점. FUTURE HORIZON, pp. 44~52.

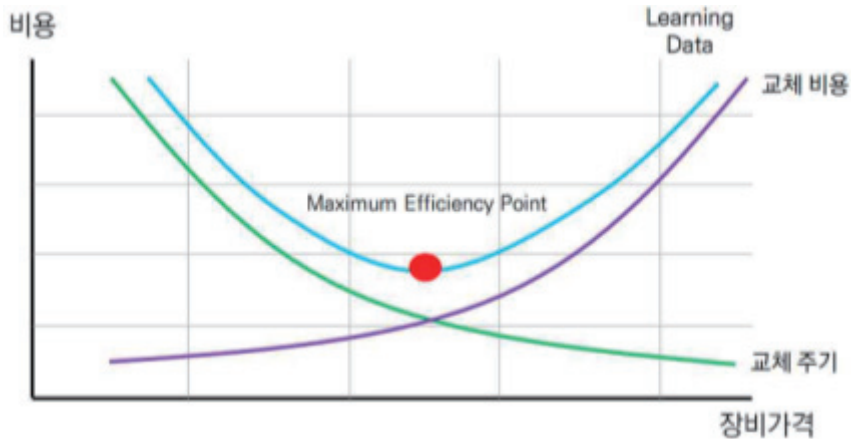
3.2.1.4. 무기체계 인공지능 과대/과소 적합에 따른 예상되는 문제점 사례

가) 무기체계의 상태 진단 오류 및 부품 교체 주기 부적절

- (1) 무기체계에 인공지능을 적용하면, 운영 효율성이나 정확성 증가 가능
- (2) 적용 분야에 따라 학습 모델의 과대/과소 적합에 따라 의도하지 않은 결과가 도출되거나 잘못 운영될 가능성이 존재⁸⁶⁾

항목	설명
배경	- 전차 및 장갑차, 자주포, 항공기, 군함 등의 전투 장비들은 다양한 부품의 집합체 - 전투력을 유지하기 위해서는 항상 각 부품들이 목표한 기능과 성능을 보장 필요
목적	- 인공지능 기술 적용, 전투 장비의 상태 진단 및 부품 교체 주기 산출⁸⁷⁾을 통한 예산 마련 - 적정 부품 교체 주기를 적용하여, 전투장비 상시 가용성 유지
적용 방식	- 전투 장비 주요 부품에 상태 진단 센서(Sensor) 부착 - 센서에서 발생하는 정보를 저장 & 분석하여 상태 진단 및 사전 부품 교체 주기 도출

- (3) 과대 적합이 적용된 경우 부품 교체 비용 절감은 가능하지만, 부품의 조기 고장 시 장비 가용성 부족에 따른 전투력 저하 가능
- (4) 과소 적합이 적용된 경우, 수시 부품 교체에 따라 비용 증가 발생
- (5) 대응 전략 : 장비 및 부품에 따라 적정 부품 교체 주기 산출 필요



〈그림 3-4〉 무기체계 상태진단에 따른 부품 교체 등 최적화 모형

86) Rahman, A. (2020). Software Defect Prediction Using Rich Contextualized Language Use Vectors (Doctoral dissertation, Queen's University (Canada)).

87) 정혜수. (2021). 한국형 무기체계 CBM 적용 및 확대를 위한 상태기반정비 발전방향 연구. 한국산학기술학회 논문지, 22(8), pp. 631~638.

나) 치명적 자율 무기(LAWs, Lethal Autonomous Weapon System) 미동작/오동작

- (1) 자율무기체계는 human-in-the-loop방식과 human-out-of-the-loop 방식으로 운용 가능⁸⁸⁾

유형	Human-in-the-loop	Human-out-of-the-loop
방식	인간 조정자가 자율무기체계의 판단을 검증하고 대응 여부에 관여⁸⁹⁾	인간 조정자가 자율무기체계의 판단과 시행에 일체 관여하지 않음
개념도		
범위	- Learning Model의 검증 - Predict 입력값, 결과의 확인 - 대응 행동 결정	- 자율무기체계 구축 및 운영 여부에만 관여 - Audit 후 AI 개선

- (2) 자율무기체계가 human-out-of-the loop 로 동작하기 위해서는 Context analysis를 위한 다수의 Sensor를 운영
- (3) Sensor의 노후화로 인해 센서에서 센싱 정보의 오류가 발생 가능
- (4) 잘못된 센싱 정보로 인해 상황을 오판하며, 의도하지 않은 자율무기체계 동작 가능

다) 전장 피아식별 오류

(1) 사례

- (가) 2003년 4월, 이라크 전쟁에서 기계적/사용자 오류에 의한 패트리엇 미사일 발사로 아군 항공기(미국/영국 항공기 각 1대씩) 격추
- (나) 2018년 9월, 시리아의 구식 레이더가 장착된 S-200를 이스라엘 F-16에 발사, 그러나 ECM(Electronic Counter Measure) 오류에 의해 근처 비행 중인 러시아 IL-2 항공기 격추

88) Firlej, M., & Taeihagh, A. (2021). Regulating human control over autonomous systems. Regulation & Governance, 15(4), pp.1071~1091.

89) Fazl Barez, "Interactive Intelligence: Human-In-The-Loop Intelligence," TheDataLab, April 22, 2019.

- (2) 전장에서는 필요 장비나 통신 회선 등의 보급이 필수적이지만 극히 제한적인 보급만 가능하고, 보급된 통신 회선에서도 그 품질 보장 불가
- (3) 전장의 군인은 기본적으로 안면 위장 및 신체 위장을 하고 각종 장비 착용, 항공기, 전차, 차량 및 포 등도 모두 위장 수행
- (4) 이에 따라 Camera에서 수집된 Image 품질 저하나 위장 효과로 인해 자율무기체계의 목표 식별 정확도가 저하되고 이로 인하여 의도하지 않는 목표 공격 가능⁹⁰⁾
- (5) 목표와 유사한 다수의 개체가 있을 때, 현재 자율무기체계는 이를 정확하게 구분 불가



- 터키 방산업체 STM은 youtube 등을 이용하여 킬러 드론 홍보
- KARGU 드론은 인간의 개입이 전혀 없이 공중을 맴돌면서 실시간으로 공격 목표물을 발견하고, 자동적으로 재빠르게 다이빙을 시작하여 몇 미터 앞에서 자폭하는 방식으로 동작
- KARGU-2는 머신러닝 학습 알고리즘을 적용, 인간의 통제 없이 목표한 타겟을 Scan한 다음 목표 발견 시 비행·자폭
- Drone의 안면 식별에 따른 신원 확인 및 살상 장면, 영화 Engel has fallen, 2019.

다) 전투원 및 비전투원 구분의 어려움 발생⁹¹⁾

- (1) 전쟁 발발시에는 열악한 보급으로 인해 민간인과 군인 모두 유사한 체형과 복장 소지품을 가져 이러한 환경에서 인공지능 기반 자율무기체계는 목표 지점에 위치한 인력의 신체적 특징, 복장과 소지품을 기준으로 전투원과 비전투원을 구별하기 어려움⁹²⁾
- (2) 전투원이 있다고 할지라도 전투가 불가능한 상황을 판단하거나 전투 의지가 없는 상황을 분석하기 어려움

90) 최은창, “컴퓨터 알고리즘은 윤리적 전투방법을 배울 수 있을까?”, MIT Technology Review, 2022년 3월 7일, <https://www.technologyreview.kr/컴퓨터-알고리즘은-윤리적-전투방법을-배울-수-있을/>

91) Russell, S. (2015). Take a stand on AI weapons. Nature, 521(7553), pp.415~416.

92) Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of artificial intelligence. Nature, 521(7553), pp.415~416.

- (3) 비전투원을 식별하였다고 하더라도 당사자의 대피나 자가 방호를 권고, 물리적 위치 이동을 강제로 하기에는 불가하므로 비전투원이 있더라도 이를 무시하고 공격 강행 가능성 증대

3.3. 소결론

3.3.1. 인공지능이란?

3.3.1.1 인공지능은 인간의 지적 능력을 기계로 구현하는 과학기술로서, ‘인간의 학습 능력과 추론 능력, 지각 능력, 자연언어의 이해 능력 등 인간의 지적 능력을 컴퓨터 프로그램으로 구현하는 과학기술’로 정의

3.3.1.2. 인공지능의 이점과 신뢰성 있는 인공지능 구현을 위한 당면 과제

- 가) 인공지능 기능 활용에 따른 장점은 복잡한 문제해결, 비즈니스 효율성 증대, 스마트한 의사결정, 비즈니스 프로세스 자동화를 이룰 수 있는 부분임
나) 그러나, 신뢰성 높은 인공지능 구현을 위해서는 다음과 같은 당면 과제가 해소되어야 함.
- (1) 데이터 거버넌스 정책은 규제 제한, 개인 정보 보호법 준수이며 이를 통해 인공지능을 구현해야 함.
 - (2) 목적에 맞는 인공지능 구현을 위한 기술적 어려움 해결방안은 많은 훈련 데이터(리소스 등)가 필수적으로 강력한 인프라가 구축되어야 함.
 - (3) 데이터 제한에 대한 해결이 필수적인 즉, 편향성 없는 데이터, 데이터의 정확성을 보장하기 위한 효과적인 데이터관리 및 품질관리가 필요

3.3.2. 인공지능 윤리에서 말하는 윤리의 개념

3.3.2.1. 인공지능에서 말하는 윤리

- 가) 인공지능 시스템이 가지고 있어야 하는 윤리적 원칙, 가치, 행동 지침을 의미
나) 인공지능은 점점 더 사회 시스템에서 중요한 역할을 수행하며 사회, 경제, 문화, 정치 등 다양한 영역에 영향을 미치게 되면서 그 사용과 개발에 대한 윤리적 고려가 더욱 중요해지고 있는 실정임.

3.3.2.2. 인공지능의 윤리원칙

- 가) 공정성과 투명성, 개인정보 보호와 프라이버시, 안전과 보안, 책임과 책임 소재, 사회적 영향과 공익, 자율성과 의사결정 투명성, 사람 중심적 가치, 지속 가능성임.
- 가) 이러한 윤리적 원칙은 인공지능 개발자, 이용자, 정부, 기업 및 기술 커뮤니티에게 인공지능의 윤리적인 사용과 발전을 촉진하며, 사회적으로 수용 가능한 방향으로 인공지능 기술을 유도해야 함.

3.3.2.3. 인공지능 윤리의 필요성

- 가) 인공지능 기술의 급속한 발전으로 제조·의료·교통·환경·교육 등 산업 전 분야에 인공지능이 활용·확산되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 인공지능 윤리 이슈가 새롭게 대두
- 나) 인공지능에 대한 법적 규제의 불충분성으로 규범적 문제에 대응하는 법적 규제가 마련되지 않았다는 점
- 다) 윤리를 통한 자율적 규제는 인공지능 과학기술이 원활하고 창의적으로 발전할 수 있도록 하면서도, 다른 한편으로 인공지능이 유발하는 규범적 문제에 적절하게 대처하기 위해서는 법적 규제보다 직접성이나 강도가 약한 윤리적 규제를 고려할 필요가 있다는 것으로 인공지능과 윤리 문제를 논의할 필요가 있음.

3.3.2.4. 인공지능 윤리에서 말하는 윤리 개념

- 가) 인공지능 윤리에서 말하는 윤리는 자유주의가 강조하는 도덕 개념으로 한정하여 이해하는 것이 적절하다고 판단
- 나) 현 시점에서 구현되는 있는 인공지능은 약한 인공지능으로 인간에 대한 관계에서 옳음을 준수하는 것으로 인간의 권리나 이익을 침해하지 않는 것이라 할 수 있음.
- 다) 따라서, 현 시점에서 인공지능에게 요구해야 하는 윤리는 그 의미 내용이 포괄적이면서도 모호한 윤리보다는 외연이 좁으면서도 그 의미 내용이 구체적인 도덕으로 보아야 함.

3.3.2.5. 윤리의 구별 개념

- 가) 인공지능 윤리는 공학윤리 가운데 인공지능을 대상으로 하여 전문적으로 특

화된 영역이라고 할 수 있으나, 인공지능 윤리는 인공지능을 설계하는 설계자에게 적용되는 윤리뿐만 아니라 인공지능 자체가 준수해야 하는 윤리, 다시 말해 인공지능 자신을 위한 윤리까지 논의 대상에 포함됨.

- 나) 로봇 윤리와 인공지능 윤리를 비교하면 인공지능은 인간의 의식, 즉 의식체계를 인공적으로 구현한 것에 해당되며, 로봇은 인간의 몸 즉, 생명체계를 인공적으로 구현한 것으로 구별되지만, 인공지능과 로봇이 결합되어 등장하는 오늘날의 상황을 고려할 때 양자를 거의 같은 의미로 사용하는 것도 문제가 없을 것으로 판단됨.

3.3.3. 인공지능 윤리의 수범 주체 및 기본 방향

3.3.3.1. 인공지능 윤리 관련자

- 가) 인공지능 윤리 제정시 수범 주체는 현시점에서 인공지능 개발자 위주의 제정이 우선시 되어야 하며, 향후 인공지능 발전에 따라서 인공지능 자체 및 이용자에 대한 윤리 기준 및 법적인 인정 부분도 고려해야 함.
- 나) 결과적으로 인공지능 윤리는 정확하게 말하면 인공지능 개발자가 준수해야 하는 윤리, 인공지능 개발자의 윤리라고 보는 것이 타당함.

3.3.3.2. 인공지능 윤리 기본 방향

가) 공리주의 측면

- (1) 공리주의적·이익형량적 사고를 명시적으로 인공지능 윤리에 적용하는 시도는 일반 대중의 지지를 받지 못하고 있으며, 이는 일반적으로 우리 인간 존재는 자신이 이익형량의 대상이 된다는 사실에 심한 거부감을 갖고 있기 때문임.
- (2) 즉, 공리주의적 윤리(명시적으로)를 인공지능 윤리로 원용하는 것은 여전히 타당하지 않음.

나) 자유주의 측면

- (1) 자유주의는 행위주체의 자유를 가장 우선시 하는 윤리를 말하며 목적론적 자유주의와 의무론적 자유주의로 구분할 수 있고, 대부분의 자유주의는 이러한 의무론적 자유주의를 수용 즉, 자유를 목적이나 공리와 무관하게 가장 우선적인 규범적 가치로서 이를 의무로 인정하는 것임.

- (2) 의무론적 자유주의 또는 의무론의 견지에서 인공지능 윤리를 설정한 것으로서 SF 소설가 아이작 아시모프(Isaac Asimov)가 제시한 로봇 3원칙이 언급되며, 이는 로봇 3원칙 우선 순위 규칙을 적용함으로써 자유주의적 성격을 유지하고 있음.
- (3) 그러나 자유주의 측면에서만 인공지능 윤리를 적용한다면 이율배반적인 상황, 서로 다른 사람의 자유 또는 권리와 이익이 상호적으로 충돌할 때 한계점에 도달함.

다) 공동체주의 측면

- (1) 공동체주의 윤리는 공동체 구성원이 모두 준수해야 하는 미덕(virtue) 또는 인륜성(Sittlichkeit)을 중시하는 윤리를 말하며, 이러한 미덕의 배후에는 좋음(the good)이 놓여 있다고 말함.
- (2) 공동체주의 윤리는 상당히 추상적이고 개별 공동체의 특성에 의존하는 점도 강해 인공지능 윤리로 적용하기에는 적절하지 않다고 판단됨. 즉, 약한 인공지능만이 기술적으로 가능한 현시점에서는 인간처럼 아주 복잡한 사고과정을 거쳐야만 판단할 수 있는 공동체주의 윤리를 인공지능 윤리로서 원용하는 것은 적절하지 않다고 말할 수 있음.

라) 기본방향

- (1) 먼저 인공지능 윤리는 자유주의 윤리를 지향해야 하며, 이는 자유주의 윤리의 규범적 의미가 가장 구체적이고 명확하기 때문임.
- (2) 자유주의 윤리를 가장 우선시 하면서도 보완적으로 공리주의 윤리를 수용할 필요가 있으며, 이는 사람과 사람 사이의 자유나 권리가 충돌하는 상황에서 서로의 이익을 최적화하기 위해서는 이익형량적 사고를 어느 정도 수용해야 함.

마) 기본 방향 설정에 따른 윤리원칙: 인간의 존엄성 존중, 권리 침해 금지, 실정법 위반 금지, 이용자의 명령 복종, 투명성, 설명 가능성, 책임

3.3.4. 인공지능 관련 이슈

3.3.3.1. 인공지능 투명성, 인공지능 공정성, 인공지능 오작동, 인공지능 보안위협, 기타 부작용 사례 등이 있음.

- ✓ 자율무기체계의 사이버 공격 대응 부재 및 부족, 전쟁 발발 가능성 증대, 자율무기체계의 목적 외 용도로 사용, 생명 경시 및 인간 존엄 훼손 등 이슈 발생 우려

3.3.3.2. 군사용 인공지능 자율무기체계의 과대적합/과소적합에 따른 문제점

- 가) 과대 적합·과소 적합은, 학습 반복 수준과 학습 데이터의 분포에 따라 그 결과가 다르게 나타나므로 국방 분야의 특징상, 일부 분야(ex. 전장, 전투 방식 등)에 대한 학습데이터를 확보하기 어렵기 때문에 인공지능 적용 분야를 고려할 필요가 있음.
- 나) 이러한 현상은 무기체계의 상태 진단 오류, 부품 교체 주기 부적절, 자율무기체계의 미작동 및 오작동, 피아식별 오류 등 발생할 수 있음.

4. 既 제정된 인공지능 윤리 기준(원칙) 분석

4.1. 주요 선진국 및 국내 사례

4.1.1. 인공지능 윤리 정책 동향

- 4.1.1.1. 세계 각국과 글로벌 기업 및 주요 국제기구들은 인공지능의 활용도 증가에 따른 산업의 진흥을 모색하는 한편 윤리적인 인공지능 실현을 위한 정책을 수립해 오고 있음.
- 4.1.1.2. 인공지능 윤리에 대한 필요성을 인식하고 해외 글로벌 기업을 선두로 인공지능 기술에 대한 원칙을 제정하기 시작한 이후 인공지능 학회를 중심으로 각종 이슈를 정리하고 윤리 및 가치를 논의하기 시작했으며 장기적인 관점에서의 원칙들을 발표하기 시작함
- 4.1.1.3. 인공지능 기술과 인공지능 기반 제품 및 서비스가 활성화되고 윤리적 이슈와 개인정보 유출과 같은 문제가 지속 제기됨에 따라 인공지능 선도 국가들은 국가 정책으로 인공지능 활용을 위한 정책을 마련하여 발표하였음
- 4.1.1.4. 최근 우리나라도 인공지능 윤리 기준 및 원칙을 발표하고 기존에 발표한 인공지능 윤리 기준이 선언적으로만 그치지 않도록 각종 가이드 라인, 개발 기준 등을 마련하고 있음
- 4.1.1.5. 군사 관련 분야에서는 미국 국방성에서 국제인도법의 원칙을 준수하는 범위내에서 군사 분야의 특성을 고려하여 국방분야에 적용할 인공지능 윤리원칙을 제시하였음

4.1.2. 주요 선진국 사례

4.1.2.1. 아실로마 (Asilomar) AI 원칙 (2017.1)⁹³⁾

- 가) 2017년 미국 캘리포니아 아실로마에서 열린 AI 컨퍼런스에서 비영리단체 Future of Life Institute는 연구 이슈, 윤리 및 가치, 장기적 이슈에 관한 23개의 AI 원칙을 발표함.
- 나) 미 캘리포니아 의회는 2018년 9월 7일 아실로마 AI 원칙을 의회결의안 (ACR-215 23 Asilomar AI Principles)으로 제정하였음.

93) Future of life institute, Asilomar AI Principles, 2017, <https://futureoflife.org/ai-principleskorean/>

〈표 4-1〉 아실로마 AI 원칙

구분		내용
1	연구목표	인공지능 연구의 목표는 방향성이 없는 지능을 개발하는 것이 아니라 인간에게 유용하고 이로운 혜택을 주는 지능을 개발하는 것이다.
2	연구이슈 연구비 지원	- 인공지능에 대한 투자에는 컴퓨터 과학, 경제, 법, 윤리 및 사회 연구 등의 어려운 질문을 포함해 유익한 이용을 보장하기 위한 연구비 지원이 수반되어야 한다. - 어떻게 미래의 인공지능 시스템을 강력하게 만들어 오작동이나 해킹 피해 없이 우리가 원하는 대로 작업을 수행하도록 할 수 있나? - 사람들의 자원과 목적을 유지하면서 자동화를 통해 우리 번영을 어떻게 성장시킬 수 있나? - 인공지능과 보조를 맞추고 인공지능과 관련된 위험을 통제하기 위해, 보다 공정하고 효율적으로 법률 시스템을 개선할 수 있는 방법은 무엇인가? - 인공지능은 어떤 가치를 갖추어야 하며, 어떤 법적 또는 윤리적인 자세를 가져야 하는가?
3	과학정책 연결	인공지능 연구자와 정책 입안자 간에 건설적이고 건전한 교류가 있어야 한다.
4	연구문화	인공지능 연구자와 개발자 간에 협력, 신뢰, 투명성의 문화가 조성되어야 한다.
5	경쟁 피하기	인공지능 시스템 개발팀들은 안전기준에 대비해 부실한 개발을 피하고자 적극적으로 협력해야 한다.
6	안전	인공지능 시스템은 작동 수명 전반에 걸쳐 안전하고 또 안전해야 하며, 적용 가능하고 실현 가능할 경우 그 안전을 검증할 수 있어야 한다.
7	윤리 투명성	인공지능 시스템이 손상을 일으킬 경우 그 이유를 확인할 수 있어야 한다.
8	사법적 투명성	사법제도 결정에 있어 자율시스템이 사용된다면, 권위 있는 인권 기구가 감사할 경우 만족스러운 설명을 제공할 수 있어야 한다.
9	가치 책임	고급 인공지능 시스템의 디자이너와 설계자는 인공지능의 사용, 오용 및 행동의 도덕적 영향에 관한 이해관계자이며, 이에 따라 그 영향을 형성하는 책임과 기회를 가진다.
10	가치관 정렬	고도로 자율적인 인공지능 시스템은 작동하는 동안 그의 목표와 행동이 인간의 가치와 일치하도록 설계되어야 한다.

구분		내용	
11	윤리 및 가치	인간의 가치	• 인공지능 시스템은 인간의 존엄성, 권리, 자유 및 문화적 다양성의 이상 에 적합하도록 설계되어 운용되어야 한다.
12		개인정보 보호	• 인공지능 시스템의 데이터를 분석 및 활용 능력의 전제하에, 사람들은 그 자신들이 생산한 데이터를 액세스, 관리 및 통제할 수 있는 권리를 가져야 한다.
13		자유와 개인정보	• 개인정보에 관한 인공지능의 쓰임이 사람들의 실제 또는 인지된 자유를 부당하게 축소해서는 안 된다.
14		공동이익	• 인공지능 기술은 최대한 많은 사람에게 혜택을 주고 힘을 실어주어야 한다.
15		공동번영	• 인공지능에 의해 이루어진 경제적 번영은 인류의 모든 혜택을 위해 널리 공유되어야 한다.
16			
17		인간의 통제력	• 인간이 선택한 목표를 달성하기 위해 인간은 의사결정을 인공지능 시스 템에 위임하는 방법 및 여부를 선택해야 한다.
		비파괴	• 고도화된 인공지능 시스템의 통제로 주어진 능력은 건강한 사회가 지향 하는 사회적 및 시정 과정을 뒤엎는 것이 아니라 그 과정을 존중하고 개 선해야 한다.
18	장기적 이슈	인공지능 무기 경쟁	• 치명적인 인공지능 무기의 군비경쟁은 피해야 한다.
19		인공지능 능력에 관한 주의	• 합의가 없으므로 향후 인공지능 능력의 상한치에 관한 굳은 전제는 피해 야 한다.
20			
21		중요성	• 고급 인공지능은 지구 생명의 역사에 심각한 변화를 가져올 수 있으므로, 그에 상응한 관심과 자원을 계획하고 관리해야 한다.
22		위협	• 인공지능 시스템이 초래하는 위협, 특히 치명적인 또는 실존적 위협에는, 예상된 영향에 맞는 계획 및 완화 노력이 뒷받침되어야 한다.
		재귀적 자기 개선	• 인공지능 시스템이 재귀적 자기 복제나 자기 개선을 통하여 빠른 수적 또 는 품질 증가를 초래한다면, 설계된 시스템은 엄격한 안전 및 통제 조치 를 받아야 한다.
23		공동의 선	• 초지능은 널리 공유되는 윤리적 이상을 위해, 그리고 몇몇 국가나 조직이 아닌 모든 인류의 이익을 위해 개발되어야 한다.

- ✓ 아실로마 AI 원칙은 인공지능 개발과 관련하여 안전, 투명성, 책임, 인간의 가치, 개인정보 보호, 공동 이익 등의 윤리적 기준을 제시하고 있음.

4.1.2.2. EU, ALTAI⁹⁴⁾ 발표 (‘20. 7.)

- 가) EU의 독립적 자문위원회인 ‘인공지능 고위 전문가 자문단’(AI HLEG⁹⁵⁾)에서 ‘신뢰할 수 있는 인공지능을 위한 자체 평가 목록’인 ALTAI를 발표함.
- 나) 총 2년에 걸쳐 ALTAI 초안을 두 차례에 걸쳐 공개하였으며 이해관계자 협의를 통해 내용을 보완함.
- 다) ALTAI 평가 목록은 자체 평가를 위한 목적과 EU가 제시하는 7대 핵심요건의 이행도구로써 활용하기 위해 개발되었음.
- 라) ALTAI는 인공지능 설계자와 개발자, 데이터 과학자, 구매조달 업무 담당자, 인공지능 시스템을 현장에서 활용할 직원들에게 권장됨.

〈표 4-2〉 EU의 ALTAI

구 분	주요 내용
기본권	• 인간의 존엄성, 비차별성, 데이터 보호, 개인정보 보호 등 기본적 영향평가를 위한 참고 문항 제시
인간의 기본권 · 자율성보장	• 인간이 최종 사용자로서 자율성을 보장
기술적 견고성 및 안정성	• 외부 공격에 대한 회복성과 보안성, 전반적인 안전, 정확성, 신뢰성, 재현성
개인정보 및 데이터 거버넌스	• 개인정보보호 강화, 데이터 관리 등에 관한 세부 사항
투명성	• 추적 가능성, 설명가능성, 소통에 대한 세부 요소
다양성, 차별금지 및 공정성	• 편견 방지 절차, 접근성, 이해관계자 참여 메커니즘
사회 및 환경적 영향	• 자연환경과 사회 전체에 미치는 영향
책임성	• 감사 가능성, 위험 관리에 관한 세부 사항

- ✓ 총 143개의 방대한 체크리스트 문항을 포함하여, 산업별·기업별 다른 특성을 가진 기업들이 자신의 상황에 맞게 활용할 수 있도록 작성함
- ✓ 자율성을 최대화하고 인공지능이 미칠 수 있는 잠재적인 영향에 대한 인식을 제고함

94) ALTAI : The Assessment List for Trustworthy Artificial Intelligence

95) HLEG: Hig-Level Expert Group on Artificial Intelligence

4.1.2.3. 영국, 데이터 윤리 프레임워크 발표 (20. 9.)

- 가) 영국의 CDDO CDDO : Central Digital and Data Office⁹⁶⁾ (중앙 디지털 및 데이터 오피스)는 공무원들의 윤리적 고려사항 이해 및 프로젝트 해결을 돕고 책임감 있는 혁신을 장려하기 위해 ‘데이터 윤리 프레임워크’(Data Ethics Framework)를 발표함.
- 나) 문서의 목적은 정부와 공공기관의 적절하고 책임성 있는 데이터 사용을 목적으로 하며, 적용 대상에는 데이터 실무자, 정책 입안자, 운영 직원 및 데이터 관련 직원 등을 비롯한 공공부문의 데이터를 직·간접적으로 활용하는 모든 인원이 포함됨.
- 다) 적용대상자는 새로운 프로젝트를 계획하고, 구현하고, 평가하는 과정 전반에 걸쳐 이 프레임워크를 참고해야하며, 데이터 수집·저장·분석·공유 등 프로세스가 변경되는 경우 프로젝트 전반에 걸쳐 프레임워크를 다시 검토해야 함.
- 라) 프레임워크는 프로젝트의 모든 측면에서 중점을 두어야 하는 가치인 ‘중요 원칙(Overarching Principles)’과 프로젝트의 다양한 절차를 안내하고 실질적인 고려사항을 제공하는 ‘구체적 조치(Specific Action)’로 구성된다.

〈표 4-3〉 영국의 데이터 윤리 프레임워크의 중요원칙

중요 원칙	주요 내용
투명성	- 투명성은 프로젝트에 대한 정보를 이해하기 쉽고 완전한 형태로 공개하며, 정보에 대한 접근성을 보장하는 것을 의미 - 프로젝트에 대한 정보가 대중에게 널리 제공될 경우 투명성을 충분히 충족한다고 평가
책임성	- 책임성은 모든 프로젝트에 대한 효과적인 감독·거버넌스 메커니즘을 갖추는 것을 의미. 공적 책임성이란 대중이 정부의 결정과 조치에 대해 효과적으로 감독·통제할 수 있음을 의미 - 장기적인 감독, 공개적인 정밀조사 메커니즘이 프로젝트 주기에 포함된 경우 책임성을 충분히 충족한다고 평가
공정성	- 프로젝트와 그 결과가 개인의 존엄성을 존중하며 공익(인권, 민주주의 가치 포함)과 일치하도록 해야 함 - 프로젝트가 공정하고 공평한 결과를 촉진하고, 무시해도 될 수준의 해로운 영향만 있으며, 국제적 인권 고려사항과 일치할 경우 공정성을 충분히 충족한다고 평가

96) CDDO : Central Digital and Data Office

〈표 4-4〉 영국의 데이터 윤리 프레임워크의 구체적 조치

중요 원칙	주요 내용
사용자 요구와 공익에 대한 이해 및 정의(9개)	• 공공 부분 데이터 프로젝트를 시작할 때 그 목적을 명확하게 설명할 수 있어야 함 • 목적에는 프로젝트가 달성하고자 하는 공공 이익과 직접적 영향을 받는 사람들의 요구가 포함되어야 함
다양한 전문가 참여(5개)	• 광범위한 기술을 보유한 다양한 전문가가 팀에 참여하는 것은 모든 데이터 또는 기술 프로젝트의 성공에 중요 • 프로젝트에 적절한 전문 지식을 가진 팀(또는 더 넓은 네트워크)의 구성원 참여 필요
법령 준수(8개)	• 데이터 관련 법령을 이해하고 있어야 하며, 확실하지 않은 경우 관련 전문가와 상의 필요
데이터 품질 및 한계 검토(9개)	• 프로젝트에 사용되는 데이터가 정확성, 대표성, 비례성, 품질 우수성을 충족하고, 한계가 설명 가능한지 확인 필요
보다 광범위한 정책 영향 평가 및 고려(7개)	• 프로젝트 결과가 책임감 있게 활용되는지를 지속적으로 평가하는 계획을 갖추어야 함 • 프로젝트 결과 및 데이터 모델이 견고한 평가 계획과 효과적인 책임 메커니즘을 통해 어떻게 사용되고 모니터링되어야 하는지 개발팀·구현팀 모두 이해할 수 있어야 함

✓ 원칙의 적용과 행동 실천을 구체적 점수로 평가하고, 원칙 중 하나라도 3점 이하의 평가를 받을 경우 문제가 있음을 명시하며, 이에 대한 팀내 또는 관련 그룹과의 논의를 통해 해결책을 찾아야 함을 강조함.

✓ 실질적으로 실무자들이 확인할 수 있는 수치적 근거를 제시할 수 있도록 하였음

4.1.2.4. OECD, 인공지능 신뢰성 확보를 위한 도구 발표 (‘21. 6.)

가) OECD는 2019 G20 디지털 경제 장관회의에서 수록된 장관 선언문에서 5가지 원칙 발표함.

〈표 4-5〉 OECD 인공지능 5원칙

원칙	내용
포용적 성장 및 지속 가능한 개발	인공지능은 포용적 성장, 지속 가능한 개발 및 웰빙을 추진함으로써 사람과 지구에 이익을 가져와야 함.
공정성 및 인간중심적 가치	- 인공지능 시스템은 법치, 인권, 민주적 가치 및 다양성을 존중하는 방식으로 설계되어야 함. - 공정하고 공정한 사회를 보장하기 위해 필요한 경우 인간 개입을 가능하게 하는 적절한 보호장치를 포함해야 함.
투명성 및 설명 가능성	사람들이 인공지능 기반 결과를 이해하고 이에 도전할 수 있도록 인공지능 시스템에 대한 투명성과 책임있는 공개가 있어야 함.
견고성과 보안성 및 안정성	인공지능 시스템은 수명주기 내내 견고하고 안전하며 안전한 방식으로 작동해야 하며, 잠재적인 위험을 지속적으로 평가하고 관리해야 함.
책임성	인공지능 시스템을 개발, 배포 또는 운영하는 조직과 개인은 위의 원칙에 따라 적절한 기능에 대해 책임을 갖음.

나) 2021년 6월 신뢰할 수 있는 인공지능 구현을 돕기 위해 ‘인공지능 신뢰성 확보를 위한 도구’(Tools for Trustworthy AI : A framework to compare implementation tools for trustworthy AI systems)을 발표함.

〈표 4-6〉 OECD AI 프레임워크 개발을 위한 분류

주요 도구	주요 내용
기술적 도구 (Technical)	인공지능 시스템의 편향성, 투명성 및 설명가능성, 견고성, 안전성 및 보안성 등 인공지능 관련 문제를 기술적 접근법으로 해결하는 것을 목표로 하는 도구·툴킷, 소프트웨어 도구, 기술 문서, 인증 및 표준, 제품 개발 도구, 기술 검증 도구 등이 포함
절차적 도구 (Procedural)	운영 또는 프로세스와 관련된 지침을 제공하기 위한 도구·가이드 라인, 거버넌스 프레임워크, 리스크 관리 도구, 부문별 행동 강령 및 단체 협약, 프로세스 인증 및 표준이 해당
교육적 도구 (Educational)	인공지능 시스템 구현에 관여하거나 영향을 받는 이해관계자의 인식, 역량을 강화하고 정보를 제공하는 것을 목적으로 하는 도구·시스템 변화대응 관리 프로세스, 포괄적 설계지침, 교육 자료 등이 포함

- ✓ 개별 도구의 품질이나 완성도는 평가하지 않았으며, 정책 입안자 또는 실무자가 필요한 도구를 신속하게 식별할 수 있도록 참고자료를 제공하는 데 중점을 두고 제작됨
- ✓ 인공지능 도구를 유형별로 분류·분석한 후, 각 도구들의 정보를 구조화하고, 목적이나 활용 맥락이 서로 다른 도구들을 비교하기 위한 프레임워크 개발함

4.1.2.5. UNESCO, 인공지능 윤리 권고안 발표 (21. 11.)

- 가) UNESCO는 2018년 초 집행위원회의 결정에 따라 과학기술윤리위원회 중심의 인공지능 윤리 관련 전문가가 보강된 확대 전문가 그룹을 운영
- 나) 확대 전문가 그룹은 2019년 3월 인공지능 윤리원칙에 대한 예비보고서 제출)
- 다) 2019년 11월 UNESCO 40차 총회에서 193개 회원국이 인공지능 윤리 권고안 제정에 대해 결의함
- 라) 2020년 3월 인공지능 윤리 권고 초안을 작성할 24인의 비상설전문가그룹(AHEG⁹⁷⁾)을 구성함
- 마) 2020년 5월 인공지능의 발전과 활용 확산에 따라 제기되는 위험과 윤리적 이슈에 대응하기 위해 ‘인공지능 윤리 권고안’(Recommendation on the Ethics of AI)을 공개함
- 바) 작성된 초안은 국제 전문가 그룹, 지역별 이해 관계자 협의회 등 의견수렴 절차를 거쳐 수정한 후 2021년 4월과 6월에 두 차례 정부간 특별위원회에서 논의를 진행함
- 사) 2021년 11월 UNESCO 41차 총회에서 회원국들의 만장일치로 채택됨

〈표 4-7〉 UNESCO 인공지능 윤리 권고

구 분	주요 내용		
목 표	- 국제법에 입각해 회원국에게 인공지능 관련 법률, 정책, 기타 도구의 제정에 관한 지침 제공을 위해 가치·행동 프레임워크를 제공 - 인공지능 시스템 수명주기 전반에 윤리적 가치 및 원칙을 정착시키기 위해 개인, 집단, 공동체, 기관 및 민간기업들에게 행동 지침 제공 - 평등의 가치, 문화적 다양성을 보호, 증진, 존중하고 현 세대와 미래 세대의 이익을 보호하며, 환경, 생물 다양성, 생태계를 보존 - 인공지능 시스템 관련 문제에 있어 이해관계자 간 다원적이고 다자적인 대화와 합의 추구 - 인공지능 분야, 이익 분배, 정보에 있어 공정한 접근을 장려		
4대 가치	- 인권 기본적 자유, 인간 존엄성의 존중 보호 증진 - 환경 및 생태계의 변형 - 다양성 및 포용성 보장 - 평화롭고 정의로우며 상호 연결된 사회 구축		
10대 원칙	<table border="0"> <tr> <td> - 비례성 및 무해성 - 안전 및 보안 - 공정성 및 비 차별성 - 지속 가능성 - 프라이버시권 및 데이터 보호 </td><td> - 인간의 감독 및 결정 - 투명성 및 설명 가능성 - 책임 및 의무 - 의식 및 지식정보 획득 및 이해 능력 - 다자적이고 조정 가능한 거버넌스 및 협력 </td></tr> </table>	- 비례성 및 무해성 - 안전 및 보안 - 공정성 및 비 차별성 - 지속 가능성 - 프라이버시권 및 데이터 보호	- 인간의 감독 및 결정 - 투명성 및 설명 가능성 - 책임 및 의무 - 의식 및 지식정보 획득 및 이해 능력 - 다자적이고 조정 가능한 거버넌스 및 협력
- 비례성 및 무해성 - 안전 및 보안 - 공정성 및 비 차별성 - 지속 가능성 - 프라이버시권 및 데이터 보호	- 인간의 감독 및 결정 - 투명성 및 설명 가능성 - 책임 및 의무 - 의식 및 지식정보 획득 및 이해 능력 - 다자적이고 조정 가능한 거버넌스 및 협력		

- ✓ UNESCO 회원국과 권고안에 언급된 모든 이해관계자들이 해당 권고안에 명시된 윤리적 가치, 원칙을 존중·증진·보호하고 권고 정책의 시행에 필요한 실현 가능한 모든 조치를 취해야 함을 강조

97) AHEG : Ad Hoc Expert Group

- 4.1.2.6. 일본, 인공지능 원칙 실천을 위한 거버넌스 가이드 라인 발표 (‘22. 1.)
- 가) 일본 경제산업성에서 인공지능 시스템 개발자, 운영자 등에 실무적 지침을 제공하기 위한 ‘인공지능 원칙 실천을 위한 거버넌스 가이드 라인’을 발표함
- 나) 이는 2019년 3월에 발표한 ‘인간 중심의 인공지능 사회 원칙’의 실현방안임
- 다) 가이드 라인은 인공지능 사회 구현 아키텍처 연구그룹(AI Social Implementation Architecture Study Group)의 ‘규제, 표준화, 가이드 라인, 감사 등 일본의 이상적인 인공지능 거버넌스 상태’와 해외 인공지능 원칙 및 규제 동향 조사에 기반하여 작성되었고, 공청회를 통해 수렴된 의견도 일부 반영한 결과물임
- 라) 일본 경제산업성은 본 가이드 라인에 법적 구속력은 없지만 인간 중심의 인공지능 사회 원칙이 준수되는 것과 마찬가지로 기업 등 사회 각 주체들의 자발적인 실천을 요청하였음.

〈표 4-8〉 일본의 거버넌스 가이드 라인

절차	주요 내용
환경 및 리스크 분석	• 기술의 긍정·부정적인영향을 적절하게 평가하고 사회적 수용도와 자사의 인공지능 숙련도에 대해 이해하는 것을 포함
목표의 설정	• 환경 및 리스크 분석에 근거하여 인공지능 거버넌스가 목표하는 의제를 설정하는 단계
인공지능시스템 설계	• 시스템을 실제로 구현하고 기존 목표에 얼마나 부합하는지를 인식
시스템의 운영	• 관리 시스템 및 개별 인공지능시스템의 운영 상황을 기록하고 공개해야하며, 공개하지 않을 시에는 그 이유를 제시할 수 있어야 함
시스템 평가	• 인공지능시스템의 작동을 검증해야 하며, 이때 외부 관계자의 의견도 포함되어야 함 • 이후 다시 환경 및 리스크에 대한 평가와 재분석을 진행함으로써 하나의 사이클을 완성하게 됨

- ✓ 일본 경제산업성은 본 가이드 라인이 살아있는 문서라는 점에 방점을 두고 있음.
- ✓ 인공지능 기술 발전이 여전히 활발하게 진행되고 있다는 점을 반영하면서, 인공지능의 개발·활용을 사회적으로 수용 가능한 수준에서 관리하기 위함임

4.1.2.7. 미국, 인공지능 권리장전 청사진 발표 (22. 10.)

- 가) 미국 백악관의 과학기술정책실(OSTP⁹⁸)은 2022년 10월 11일 ‘인공지능 권리장전(AI Bill of Rights) 청사진’을 공개함.
- 나) 인공지능 기술의 진전이 미국의 기본 원칙인 시민권과 민주주의적 가치를 훼손하지 않도록 보호하고자, 바이든 대통령의 비전을 권리장전 형태로 구체화한 것임.
- 다) 권리장전은 민주적 가치, 시민권, 자유, 프라이버시를 보호하는 자동화 시스템을 구축하고 배포하기 위한 다섯가지 기본원칙을 제시함.
- 라) 원칙은 안전하고 효과적인 시스템, 알고리즘에 의한 차별 방지, 데이터 프라이버시, 고지 및 설명, 인적대안과 고려사항 및 대체 등으로 구성됨

〈표 4-9〉 미국의 인공지능 권리장전 청사진

기본 원칙	주요 내용
안전효과적시스템	• 누구나 안전하지 않거나 비효율적인 시스템으로부터 보호받아야 한다.
차별 방지	• 누구도 알고리즘·시스템으로 인해 차별 받아서는 안되며, 시스템은 공평한 방식으로 설계되고 사용되어야 한다.
데이터 프라이버시	• 누구나 기본 제공된 보호기능을 통해 악의적인 데이터 관행으로부터 보호받아야 하며, 본인의 데이터가 어떻게 사용되는지에 대한 인지 능력을 가져야 한다.
고지설명	• 누구나 자동화된 시스템이 사용되고 있다는 사실을 알고, 개인에 영향을 미치는 결과에 자동화된 시스템이 기여하는 방식과 이유를 이해할 수 있어야 한다.
인적 대안대체	• 필요한 경우 누구나 선택을 철회할 수 있어야 하며, 직면한 문제를 신속하게 고려하고 해결할 수 있는 사람과 연락할 수 있어야 한다.

- ✓ 인공지능 권리장전 청사진은 프레임워크로서 기존 미국 내에서 제시되고 있는 인공지능 사용에 관련된 원칙을 전반적으로 정리한 것임.
- ✓ 단순히 원칙만을 제시한 것이 아니라 해당 원칙이 현실에서 실천으로 어떻게 이행될 수 있는지, 미국 연방정부, 국립연구소 등에서 실시하고 있는 각종 연구나 리스크 관리 프레임워크, 관련 표준에 대한 지침과 더불어 민간 사례에 대한 분석을 제공한다는 점에서 의미가 있음.

98) OSTP : Office of Science and Technology Policy

4.1.2.8. 미국 국방성, 인공지능 윤리원칙 발표 ('20. 2.)⁹⁹⁾

- 가) 미 국방성은 2018년에 국방 인공지능 전략을 발표하였음. 주된 내용으로 ① 핵심임무를 처리할 수 있는 인공지능 능력 제공, ② 분산적인 개발과 시험을 위한 공통기반을 통해 DoD 내 인공지능 영향력 확대, ③ 인공지능 선도 인력 양성, ④ 산·학 및 국제동맹들과의 협력, ⑤ 인공지능 군사윤리 및 안전성 선도를 제시 등이다.¹⁰⁰⁾
- 나) 미 국방성은 국방 인공지능 전략 발표 후 국방혁신위원회(DIB¹⁰¹⁾)를 통해 전투원과 비전투원 모두에게 적용될 수 있는 인공지능 사용 원칙을 논의하고 구체적 윤리원칙을 수립하였음.
- 다) 협력 국가들의 대표자가 포함된 윤리원칙에 대해 2번의 공청회, 3번의 자문 회의, 2번의 고위실무회의, 국방부 실무그룹의 13번의 비공식 회의를 통해 2020년 2월 인공지능 윤리원칙에 대한 5가지 원칙과 12가지 권고사항을 제시함
- 라) 미 국방성의 인공지능 윤리원칙은 미국 연방헌법, 연방법률 중 국방 관련 법령, 국제법, 특히 국제인도법(전쟁법)에 부합하는 원칙으로 수립함
- 마) 국방혁신위원회는 국방성의 5가지 원칙을 적용하고 실현하기 위해 필요한 12가지 사항에 대해 권고사항도 제시함.

〈표 4-10〉 미국 국방성 인공지능 윤리원칙

기본 원칙	주요 내용
책임성	• 적절한 수준의 주의와 판단력을 지속적으로 행사 (개발, 배치, 사용에 대한 책임도 유지)
공평성	• 인공지능성능에 있어 비의도적인 편향이 최소화되도록 신중하게 조치
추적가능성	• 관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발
신뢰성	• 인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제가능성	• 의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

99) <https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>

100) 한국국방연구원, 군사적 인공지능의 법적(윤리적) 문제에 대한 검토, 2020.4.27. pp.19~20

101) DIB : Defense Innovation Board

〈표 4-11〉 미국 국방성 인공지능 윤리원칙 권고사항

권고 사항	주요 내용
공적인 국방성 조직을 통한 공식화	합동인공지능센터는 국방부 장관에게 인공지능 윤리원칙의 지속성을 보장하기 위해 적절한 공보와 정책서의 발행을 건의
고위급 감독 위원회 설립	국방부 장관은 인공지능 윤리원칙에 따라 국방성 인공지능계획이 시행되고, 인공지능전략의 집행과 감독이 이루어지도록 보장하기 위해 고위급 위원회를 설립
인공지능 기술 분야의 성장 및 육성	연구기술차관보는 적절한 기술실무과정 등을 통해 인공지능 기술분야의 발전과 성숙화를 위한 지원을 제공
관련자의 교육과 훈련프로그램 향상	개별 국방기관은 인공지능과 관련된 기술과 지식을 보유하고 있는 국방 인력의 교육 및 훈련 프로그램을 제공
인공지능의 새로운 보안 관련 연구투자	정책차관보는 사이버보안, 양자컴퓨터, 정보작전, 생물기술 등이 인공지능과 결합하는 경우 가질 억제력과 경쟁력 확보를 위해 투자
재현력 향상을 위한 연구투자	연구기술차관보는 복잡한 인공지능이 어떻게 작동되는지 이해를 돕기 위해 재현력 향상을 위한 투자를 시행
신뢰기준의 정립	연구기술차관보는 인공지능의 실행을 측정하기 위해 필요한 최상의 기준을 정립하도록 노력
인공지능 시험평가 기술의 강화	현재의 시험평가, 검증, 인전 절차를 활용 및 발전시키고 필요하다면 인공지능 시스템을 위한 새로운 구조 마련
위기관리 방법의 발전	합동인공지능센터는 윤리, 안정성, 법적 위험의 고려를 기반으로 하여 저 위험, 고위험 기술을 구분하여 활용하는 국방 분류방법을 마련
인공지능 윤리원칙의 적절한 이행보장	합동인공지능센터는 국방수권법 조항 238에 의해 요구되는 통제와 감독의 일환으로 윤리원칙의 적절한 이해여부를 조사
인공지능 윤리이행의 이해에 관한 연구 확대	연구기술차관보는 안전, 보안, 강력한 기능에 관하여 다수의 대학 연구진이 참여하는 단체를 조직
인공지능의 안전, 보안, 강력성에 대한 연중 컨퍼런스 개최	합동인공지능센터는 다양한 계층이 참여하는 인공지능의 안전, 보안, 강력한 기능을 검토하는 컨퍼런스를 매년 개최

4.1.3. 주요 선진국 사례 범주별 분류

4.1.3.1. 제반 원칙 정리

범주	책임성	신뢰성 (안전성)	공평성	투명성 (추적, 설명)	데이터 관리	통제 가능성	인권보장 / 인간존중	프라이 버시 (개인정보)	기 타
아실로마 AI원칙	○	○		○		○	○	○	가치관정렬, 공동이익, 공동변영, 비파괴, 인공지능 무기경쟁
EU (ALTAI)	○	○	○	○	○			○	기본권, 자율성, 사회 및 환경적 영향
영국 (DEF)	○		○	○					
OECD	○	○	○	○			○	○	포용적 성장 및 지속 가능한 개발
UNESCO	○	○	○	○	○			○	비례성 및 무해성, 비차별성, 지속가능성, 인간의 감독 및 결정, 의식 및 지식정보 획득 및 이해 능력, 다자적이고 조정 가능한 거버넌스 협력
미 국방성	○	○	○	○		○			

4.1.3.2. 범주별 분류 결과

- 가) 대부분 기관에서 책임성과 투명성에 대한 원칙은 모두 포함하여 관리하고 있음.
- 나) 신뢰성과 공평성이 중요한 핵심 요소로 관리되고 있음
- 다) 통제가능성은 주요 국가에서는 채택하지 않으나 미 국방성에서는 인공지능 원칙 항목으로 선정하여 관리 중임.

4.2. 정부 부처 및 국내.외 주요 기업 사례

4.2.1. 국내 정부부처 사례

4.2.1.1. 관계부처 합동, 사람이 중심이 되는 인공지능 윤리 기준 (’20. 12.)

- 가) 과학기술정보통신부는 대통령 직속 4차산업혁명위원회 전체회의에서 사람이 중심이 되는 ‘인공지능 윤리 기준’을 심의.의결함.
- 나) 윤리적 인공지능을 실현하기 위해 정부.공공기관, 기업, 이용자 등 모든 사회구성원이 인공지능 개발에서 활용의 전 단계에서 함께 지켜야 할 주요 원칙과 핵심 요건을 제시하는 기준임
- 다) 주요 전문가들의 자문과 의견수렴 과정을 거치며 윤리철학의 이론적 논의와 연계하여 ‘인간성을 위한 인공지능(AI for Humanity)’을 목표로 하는 윤리 기준을 마련함.
- 라) 인공지능 주요 원칙과 핵심 요건으로 ‘인간성(Humanity)’을 최고 가치로 내세우며, 그에 맞춰 인간 존엄성 원칙 . 사회 공공선 원칙 . 기술의 합목적성 원칙으로 구성된 3대 기본원칙과 인권 보장, 프라이버시 보호, 다양성 존중, 침해 금지, 공공성, 연대성, 데이터 관리, 책임성, 안정성, 투명성의 10대 핵심요건을 제시함



〈그림 4-1〉 관계부처 합동(과학기술정보통신부), 사람이 중심이 되는 인공지능 윤리 기준

〈표 4-12〉 관계부처 합동(과기정통부), 사람이 중심이 되는 인공지능 윤리 기준 핵심요건

핵심 요건	주요 내용
인권 보장	• 인간의 권리와 자유 침해 금지
프라이버시 보호	• 프라이버시 보호 및 개인정보 오용 최소화
다양성 존중	• 편향 및 차별을 최소화하고 모두에게 공정하게 적용 • 취약계층에 대한 접근성 확보
침해 금지	• 해악금지 • 인공지능이 초래하는 위험 예방
공공성	• 개인의 행복 및 사회 공공성을 위해 활용
연대성	• 다양한 집단의 연대 및 참여 보장
데이터 관리	• 데이터 편향 및 데이터 품질 관리
책임성	• 개발 및 활용 과정의 책임주체 설정
안전성	• 잠재적 위험 방지 및 안전 보장 • 시스템 제어 가능성 확보
투명성	• 인공지능 활용 상황에 적합한 수준의 투명성·설명 가능성 확보

4.2.1.2. 국토교통부, 자율주행차 윤리 가이드 라인('20. 12.)

- 가) 국토교통부는 '사이버보안 가이드 라인', '레벨4 제작·안전 가이드 라인'과 함께 '자율주행차 윤리 가이드 라인'을 발표함. 이는 법 체계상 제도화하기 어려운 윤리적 방향성을 제시함으로써 국민들이 수용할 수 있는 안전한 자율주행차 도입 환경을 조성하기 위함임
- 나) 본 가이드 라인의 적용대상은 레벨4 이상 완전자율주행시스템이 장착된 자동차를 설계·제작·관리 운행하려는 자와 소비자가 대상임.
- 다) 자율주행차가 인명 보호를 최우선하도록 설계·제작되어야한다는 원칙을 중심으로 기본 가치와 행위 원칙, 그리고 자율주행차와 관련된 각 주체들이 준수해야하는 윤리원칙을 제시함

4.2.1.3. 개인정보보호위원회, 인공지능 개인정보보호 자율점검표('21. 5.)

- 가) 개인정보보호위원회는 인공지능 서비스의 개발과 운영 시 발생할 수 있는 개인정보 침해를 예방하기 위한 안내서로 「인공지능 개인정보보호 자율점검표(개발자·운영자용)」를 2021년 5월 31일 공개하였음.

- 나) 인공지능 설계, 개발·운영 과정에서 개인정보를 안전하게 처리하기 위해 지켜야 할 「개인정보보호법」 상 주요 의무·권장사항을 단계별로 자율점검할 수 있도록 정리한 안내서로, 대규모 데이터의 처리, 복잡성·불투명성, 자동화·불확실성 등 인공지능의 개인정보 처리 특성을 고려하였음.
- 다) 개인정보보호위원회는 점검표에 「개인정보보호법」, 과학기술정보통신부의 「인공지능 윤리 기준」, 국제적으로 통용되는 「개인정보보호 중심설계 원칙」 등을 반영하여 ‘인공지능 관련 개인정보보호 6대 원칙’과 이를 기반으로 단계별로 점검해야 할 16개 항목, 54개 확인사항을 담았음.

4.2.1.4. 방송통신위원회, 인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙(21. 6.)

- 가) 방송통신위원회는 2021년 6월 30일, 인공지능 기반 추천 서비스의 투명성과 다원성을 확보하는 취지에서 서비스 제공자에게 권고하는 자율규범으로서 「인공지능 기반 미디어 추천 서비스 이용자 보호 기본원칙」을 발표하였음.
- 나) 해당 원칙은 미디어 콘텐츠가 제공되고 이용되는 디지털 플랫폼 영역에 초점을 두고 미디어 이용자의 권익 보호를 위해 서비스 제공자가 실현해야 할 3대 핵심 원칙과 실천적 방안으로서 5대 실행 원칙을 제시하였음.

4.2.1.5. 국가인권위원회, 인공지능 개발과 활용에 관한 인권 가이드 라인(22. 5.)

- 가) 국가인권위원회는 2022년 5월 11일 인공지능 개발과 활용 과정에서 발생할 수 있는 인권침해와 차별을 방지하기 위하여 「인공지능 개발과 활용에 관한 인권 가이드 라인」을 마련하였음.
- 나) 인공지능을 개발할 때부터 개인의 삶과 사회적 공익에 기여할 수 있도록 개발하고 인간의 존엄성과 자기결정권의 보장 및 차별받지 않을 권리 등 인권에 기반하여 활용하며 인공지능 기술로 발생할 수 있는 인권침해의 예방 및 권리구제 절차를 마련하는 것이 중요하다는 점에서 국가인권위원회는 「국가인권위원회법」 제25조 제1항3)에 따라 인권적 관점에서 인공지능의 개발과 활용 시 준수해야 할 내용을 검토하였음.

4.2.1.6. 교육부, 교육분야 인공지능 윤리원칙(22. 8.)

- 가) 2022년 8월 교육부는 교육분야 인공지능 개발자와 교육 당사자들이 함께 준수해야 할 「교육분야 인공지능 윤리원칙」을 확정·발표하였음.

- 나) 해당 윤리원칙은 1월 27일 시안이 발표된 이후 공청회와 전문가 간담회, 국제 의견 조회 등 광범위한 의견 수렴을 거쳐 마련되었으며 교육계와 관련 산업계가 준수해야 할 원칙과 실천과제를 담고 있음.
- 다) 교육부 ‘교육분야 인공지능 윤리원칙’은 교육이 궁극적으로 사람의 성장을 지원한다는 가치에 따라 “사람의 성장을 지원하는 인공지능”을 대원칙으로 설정하고, 교육기관 및 교육활동을 지원하는 행정기관에서 활용되는 인공지능이 지켜야할 10개의 세부원칙을 제시하였음.

4.2.1.7, 국내 정부부처 사례 범주별 분류

범주	책임성	신뢰성 (안전성)	공평성 (공공성)	투명성 (추적, 설명)	데이터 관리	통제 가능성	인권보장 / 인간존중	프라이 버시 (개인정보)	기 타
관계부처 합동	○	○	○	○	○		○	○	다양성 존중, 침해금지, 연대성,
국토교통부	○	○		○				○	
개인정보보 호위원회	○	○	○	○					적법성, 참여성,
방송통신 위원회	○		○	○					
국가인권 위원회				○			○		자기결정권의 보장, 차별금지, 인공지능 인권영향평가 시행, 위험도 등급 및 관련 법 제도 마련
교육부		○	○	○	○			○	인간 성장의 잠재성, 학습자의 주도성과 다양성 보장, 교수자의 전문성 보장, 교육당사자의 관계 유지, 사회 공공성 증진

- 가) 국내 대부분의 정부 부처는 투명성을 핵심 원칙으로 관리하고 있음.
- 나) 주요 선진국 사례와 유사하게 책임성을 강조되고 있음
- 다) 신뢰성, 공평성, 프라이버시 등을 주요 인공지능 원칙으로 지정하고 있음

4.2.2. 해외 글로벌 기업 사례

4.2.2.1. Microsoft, 인공지능 원칙 발표 (‘16)¹⁰²⁾

가) 마이크로소프트는 사람들을 최우선으로 윤리적 원칙에 의한 책임 있는 인공지능 기술을 위한 6개 원칙을 마련함

〈표 4-13〉 Microsoft 인공지능 원칙

구분	주요 내용
공정	• 시스템은 모든 사람을 공정하게 대우해야 함.
신뢰성 및 안전성	• 인공지능 시스템은 안전하고 확실하게 수행해야 함.
개인정보보호/보안	• 인공지능 시스템은 안전해야 하며 개인정보를 존중해야 함.
포괄성	• 인공지능 시스템은 모든 사람에게 권한을 부여하고 사람들을 참여시켜야 함.
투명성	• 인공지능 시스템은 이해할 수 있어야 함.
책임	• 사람들은 인공지능 시스템에 책임을 져야 함.

4.2.2.2. Google, 인공지능 원칙 발표(18. 6.)¹⁰³⁾

가) 구글은 2017년 5월 미국 국방부와 군사작전에 사용될 미군 무인기로 촬영한 감시 자료를 분석하는 프로젝트 메이븐(Project Maven)을 발표했으나, 구글 직원들은 군사작전에 구글의 인공지능 기술이 사용되는 것은 비윤리적이며, 군사적 정치적으로 사용된다는 것 자체에 대해 강한 우려를 나타냄.

나) 이에 따라 2018년 6월 구글은 책임감 있게 기술을 개발하고, 추구하지 않을 특정 응용분야를 확립하겠다는 약속으로 인공지능 개발 시 준수해야 할 7개 목표를 발표함

다) 또한, 아래와 같은 영역에는 인공지능의 설계나 배포를 하지 않는다고 선언함.

- (1) 전반적인 피해를 유발하거나 발생할 가능성이 있는 기술
- (2) 무기 또는 주된 목적 또는 이행이 사람에게 상해를 유발하거나 직접적으로 촉진하는 기타 기술
- (3) 국제적으로 인정되는 규범을 위반하는 감시를 위해 정보를 수집하거나 사용하는 기술

102) <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6>

103) <https://ai.google/principles>

(4) 그 목적이 국제법과 인권의 널리 인정되는 원칙에 위배되는 기술

〈표 4-14〉 구글의 인공지능 원칙

구분	주요 내용
사회에 유익한 인공지능	• 인공지능을 사용하여 고품질의 정확한 정보를 쉽게 이용할 수 있도록 노력하면서, 사업을 영위하는 국가의 문화적, 사회적, 법적 규범을 지속적으로 존중할 것
편견이 없는 공평한 인공지능	• 인종, 민족, 성별, 국적, 소득, 성적지향, 능력, 정치적 또는 종교적 신념과 같은 민감한 특성과 관련된 사람들에게 부당한 영향을 피하기 위해 노력할 것
안전한 인공지능	• 피해의 위험을 초래하는 의도하지 않은 결과를 피하기 위해 강력한 안전 및 보안 절차를 지속적으로 개발하고 적용할 것
사람이 제어하는 인공지능	• 피드백, 관련 설명 및 호소에 대한 적절한 기회를 제공하는 인공지능 시스템을 설계할 것
개인정보 원칙을 지키는 인공지능	• 통지 및 동의 기회를 제공하고, 개인정보보호 장치를 갖춘 아키텍처를 장려하고, 데이터 사용에 대한 적절한 투명성과 통제권을 제공할 것
과학적으로 우수한 인공지능	• 다양한 이해 관계자들과 협력하여 이 분야에서 사려 깊은 리더십을 증진하고 과학적으로 엄격하고 다분야적인 접근 방식을 그릴 것
악용의 소지가 없는 인공지능	• 잠재적으로 유해하거나 학대적인 응용 프로그램을 제한하기 위해 노력할 것

4.2.2.3. IBM, 인공지능을 위한 일상 윤리 (21. 8.)

가) 인공지능의 시작 단계부터 디자인 및 개발 프로세스를 포함하는 시스템 디자이너와 개발자를 위한 인공지능 윤리 실용 가이드를 제공하여 다섯가지 주요 윤리적 영역과 각 영역별 권장 조치를 제시함

〈표 4-15〉 IBM 인공지능을 위한 일상 윤리

구분	주요 내용
책임	• 인공지능 디자이너와 개발자는 인공지능 디자인, 개발, 의사결정 과정 및 결과를 고려해야 할 책임이 있음
가치 부합	• 인공지능을 설계할 때 사용자 그룹의 규범과 가치를 염두에 두고 이에 부합하도록 해야함
설명 가능성	• 인공지능을 설계할 때 사람이 인공지능의 의사결정 과정을 쉽게 인지, 감지 및 이해할 수 있도록 해야함
공정성	• 인공지능은 편향을 최소화하고 포용적 표현을 장려하도록 설계해야 함
사용자 데이터 권한	• 인공지능은 사용자 데이터를 보호하고 액세스 및 사용에 대한 사용자의 권한을 보존하도록 설계되어야 함

4.2.2.4. Meta, 책임성 있는 AI의 5가지 기준 발표 ('21. 6.)

- 가) Meta(前 Facebook)는 인공지능이 가지는 프라이버시, 공정성, 책임성 및 투명성과 같은 문제를 직면하고, 협력적인 방식으로 문제를 해결하고자 2021년 6월 'Facebook's Five Pillars of Responsible AI'를 발표하였음.
- 나) 책임성 있는 인공지능을 위한 5가지 기준으로 ① 개인정보 보호 및 보안, ② 공정성 및 포괄성, ③ 견고성 및 안전성, ④ 투명성 및 통제, ⑤ 책임 및 거버넌스를 제시하였음.
- 다) 발표문에서 Meta AI는 인공지능이 상대적으로 새로운 분야이기 때문에 인공지능 거버넌스와 잠재적인 부정적 영향을 평가하기 위한 기준과 프로세스가 아직 명확하게 정의되지 않았다고 밝히고 있음.
- 라) 기술 산업, 인공지능 연구 커뮤니티, 정책 입안자 등은 인공지능 영향 평가가 명확하고 합리적인 기준으로 만들어질 수 있도록 협력해야 함을 강조하였음.
- 마) 영향평가를 통해 새로운 인공지능 기반 솔루션을 구축함과 동시에 잠재적인 부정적 인공지능 영향을 식별하고 해결할 수 있다는 설명을 덧붙임.

〈표 4-16〉 Meta의 책임성 있는 AI 5가지 기준

기본 원칙	주요 내용
개인정보 보호 및 보안	• 개인정보보호 및 위험 식별·완화
공정성 및 포괄성	• 잠재적 통계적 편향 탐지
견고성 및 안전	• 의도한 대로 동작하도록 설계 및 테스트
투명성 및 통제	• 책임있는 인공지능 • 투명하고 설명가능해야 하며, 사용자의 통제권에 대해 이해를 지원
책임 및 거버넌스	• 인공지능이 내리는 의사결정에 대한 책임 보장 및 신뢰할 수 있는 프로세스 구축

- ✓ 기업 스스로 시장과 산업의 방향성을 제안하고 질서를 구축하는 것이 기술발전에 따른 리스크를 최소화하는 데 있어 규제적 접근보다 효율적이고 합리적일 수 있다는 점에서 긍정적 평가가 가능함.

4.2.2.5. CISCO, 책임성 있는 인공지능 원칙 발표 ('22. 1.)

- 가) Cisco는 2022년 1월, 모든 사람의 인권을 존중하고 유지하는 것을 목표로 투명성, 공정성, 책임성, 개인정보 보호, 보안 및 신뢰성 등을 포함하는 「책임성 있는 인공지능을 위한 원칙」을 발표하였음.

- 나) 해당 원칙은 인공지능 기술의 거버넌스에 직접 적용 가능하도록 고안되었음.
 다) 각 원칙에는 구체적인 행동강령이 포함되어있으며, 고객이 개발 과정에 지속적으로 참여하고 피드백을 줄 수 있도록 지원하는 등의 내용을 담고 있음.

〈표 4-17〉 CISCO의 인공지능 원칙

기본 원칙	주요 내용
투명성	• 인공지능 기술의 투명성을 일관되고 명확하게 제공
공정성	• 유해한 편견을 식별 및 해결
책임성	• 인공지능 생애 전 주기에 개인정보 보호, 보안 항상 고려
개인정보보호	• 개인정보 처리 시 항상 개인정보 법률 및 표준 준수
정보보안	• 복원력 및 신뢰성 강화
안정성	• 의도한 대로 작동하는 책임있는 인공지능 설계 및 테스트

- ✓ Cisco의 원칙은 해당 산업의 특성상 기술의 보안에 초점을 두며, 인공지능의 안전성, 정보 보안, 개인정보 보호 등을 강조하고 있음.

4.2.2.6. 해외 글로벌기업 사례 범주별 분류

범주	책임성	신뢰성 (안전성)	공평성 (공공성)	투명성 (추적, 설명)	데이터 관리	통제 가능성	인권보장 / 인간존중	프라이 버시 (개인정보)	기 타
Microsoft	○	○	○	○				○	포괄성
Google		○	○			○		○	사회에 유익한 AI, 과학적으로 우수한 AI, 악용의 소지가 없는 AI
IBM	○			○	○				가치 부합
Meta	○	○	○	○				○	
CISCO	○	○	○	○				○	정보보안

- 가) 해외 글로벌 기업은 책임성, 신뢰성, 공평성, 투명성 등을 핵심 항목으로 관리하고 있음
 나) 프라이버시와 정보보안 항목도 주요한 인공지능 원칙으로 지정함

4.2.3. 국내 주요 기업 사례

4.2.3.1. 카카오, 카카오 알고리즘 윤리헌장(18. 1.)

- 가) 카카오는 2017년 3월 인터넷 기업 최초로 뉴스 배열 알고리즘을 공개하면서 해당 알고리즘을 설명하였음. 이는 진화된 뉴스 알고리즘 공개를 통한 알고리즘 설명에 대한 사회적 책무를 이행하겠다는 회사의 의지를 반영함
- 나) 그 후 인공지능 윤리 의식 정비와 알고리즘 윤리원칙 제정을 위한 준비에 착수하였으며, 카카오 내부 전체 직원들의 의견 수렴 및 적합성 검증을 진행하였음
- 다) 2018년 1월 카카오 알고리즘 윤리헌장을 발표하였고, 2019년 8월에 ‘기술의 포용성’ 신규 조항 추가한 개정본을 발표함. 2020년 7월에 2차 개정을 통해 ‘아동과 청소년에 대한 보호’ 조항을 신규로 추가함.

4.2.3.2. 네이버, 인공지능 윤리 준칙 (21. 2.)

- 가) 네이버는 서울대학교와 함께 「인공지능 윤리 준칙」을 만들어 발표했음.
- 나) 네이버 「인공지능 윤리 준칙」은 ① 사람을 위한 인공지능 개발, ② 다양성의 존중, ③ 합리적인 설명과 편리성의 조화, ④ 안전을 고려한 서비스 설계, ⑤ 프라이버시 보호와 정보 보안 등 총 5가지 원칙으로 구성되어 있으며, 네이버는 학계와 협업을 통해 전문가들의 의견을 반영하고 자사의 기업 철학도 함께 고려해 윤리 준칙에 반영했음.
- 다) 네이버는 윤리 준칙을 통해 인공지능이 사람을 위한 기술이며, 우리의 삶을 편리하게 만들어줄 수 있지만 완벽하진 않기 때문에 꾸준히 개선하겠다는 점을 강조하는 한편, 향후 인공지능 윤리 준칙 실천 사례 추적 및 개선을 위한 학계와의 지속적 협력 계획도 밝히며 윤리 준칙을 계속 보완할 것을 시사하였음

4.2.3.3. SK텔레콤, SKT 인공지능 추구 가치 제정(21. 5.)

- 가) 2021년 5월 13일 SK텔레콤은 「사람 중심의 인공지능」을 핵심이념으로 하는 인공지능 윤리 가치 원칙을 발표했음.
- 나) SK텔레콤은 ‘인공지능 기업(AI Company)으로의 혁신’을 목표로 전사 차원의 인공지능 원칙 정립을 추진해왔으며 구성원 워크숍, 임원 인터뷰 등을 통해 인공지능을 개발하면서 주요하게 고려했던 원칙을 확인하고, 국내외 인공지능 윤리 기준에 대한 분석, 전문가들의 의견 수렴을 진행하였고, 이러한 과정을 통해 ‘인공지능 추구 가치 원칙’을 정립했음.

4.2.3.4. LG, 인공지능 윤리원칙(22. 8.)

- 가) 2022년 8월 23일 LG는 신뢰할 수 있는 인공지능을 개발하기 위해 필요한 원칙인 「인공지능 윤리원칙」을 발표했다.
- 나) 인공지능 윤리원칙은 인공지능을 개발하는 LG의 모든 구성원이 지켜야 하는 행동과 가치 판단의 기준이 되는 원칙으로 밝히며, 구체적으로 인간존중(Humanity), 공정성(Fairness), 안전성(Safety), 책임성(Accountability), 투명성(Transparency) 등 5가지 핵심 가치를 인공지능 윤리원칙으로 강조했다.

4.2.3.5. 국내 주요기업 사례 범주별 분류

범주	책임성	신뢰성 (안전성)	공평성 (공공성)	투명성 (추적, 설명)	데이터 관리	통제 가능성	인권보장 / 인간존중	프라이 버시 (개인정보)	기 타
카카오					○			○	카카오 알고리즘의 기본원칙, 차별에 대한 경계, 알고리즘의 독립성, 알고리즘에 대한 설명, 기술의 포용성, 아동과 청소년에 대한 보호
삼성 전자	○		○	○					
네이버		○		○			○	○	다양성의 존중
SK 텔레콤		○	○	○				○	사회적 가치, 무해성, 지속혁신
LG	○	○	○	○			○		

가) 국내 주요 기업은 해외 글로벌 기업과 유사하게 인공지능 원칙을 제정함

나) 개인정보와 관련된 프라이버시 항목도 주요 항목으로 관리함

다) 데이터를 활용하는 기업에서는 데이터 관리를 주요 인공지능 원칙으로 지정함

4.3. 시사점

4.3.1. 주요 선진국 사례 시사점

4.3.1.1. 인공지능 윤리적 사용을 위한 다양한 방안 제시 중

- 가) 가이드 라인과 프레임워크를 개발하여 인공지능의 윤리적 개발과 활용을 자율적으로 규율할 수 있는 체계를 수립
- 나) 인공지능 윤리 기준 및 원칙이 실질적으로 바람직한 방향으로 진행될 수 있도록 가이드 라인과 프레임워크 제시함

4.3.1.2. 인공지능 윤리 기준 및 원칙의 실질적 이행방안 모색 중

- 가) 법률적 규제를 통하여 강제력을 부여하려는 동향도 보임
- 나) 많은 국가들은 자율적 규제를 모색함과 동시에 해당 규범의 준수를 강제하는 법률을 통해 인공지능 기술의 윤리적 활용을 규율하고자 하는 움직임을 보임¹⁰⁴⁾

4.3.1.3. 인공지능 윤리의 자발적 참여 유도

- 가) 인공지능 개발자 혹은 사용자에게 인공지능 윤리 규범을 준수하도록 강제할 수 없으며 자발적으로 준수하지 않을 경우 해당 규범은 아무런 의미를 가지지 못함
- 나) 해당 규범에 대한 자의적 해석 문제가 발생하여 실질적인 통제가 어려워질 수 있음.

4.3.2. 해외 글로벌 기업 사례 시사점

4.3.2.1. 인공지능 기술 발전에 따른 부작용 우려

- 가) 인공지능 기술과 서비스가 일상화되고 보편화되는 가운데 잠복해 있던 윤리 훼손 이슈에 대해 관심도가 높아졌음.
- 나) 인공지능을 더욱 적극적으로 개발하고 활용범위를 최대한 확대하고자 할수록 인공지능 활용을 통한 이익의 확보와 인공지능에 대한 국민의 보호를 균형적으로 이루어야 함.

104) 박혜성, 김법연, 권현영 (2021), 인공지능 통제를 위한 규제의 동향과 시사점, 정보법학, 25권, 1-39

4.3.2.2. 인공지능 윤리의 확산

- 가) 인공지능 서비스로 인한 부작용이나 위험을 최소화하기 위해서는 기업의 사회적 책임에 수반되는 윤리의식의 선행이 필요하다는 공감대가 형성됨
- 나) 인공지능의 결함을 해결하지 못하고 인공지능 서비스를 시장에 출시할 경우 엄청난 법적, 재무적 위험에 직면한다는 현실을 직시하게 되었음.
- 다) 이러한 법적, 재무적 손실과 더불어 기업은 기존에 쌓아왔던 명성과 같은 유·무형 손실이 가중될 수 있다는 것을 인지함.

4.3.3. 국내 정부 부처 사례 시사점

4.3.3.1. 인공지능 윤리 기준 제정을 위한 정부의 적절한 개입 필요

- 가) 인공지능 윤리원칙 준수를 위한 기업의 자율적 노력도 필요하지만 관련 정보를 교환하고 자율적 규정을 마련할 수 있도록 정부의 가이드 라인 제정이 필요함
- 나) 인공지능 기술이 적용되는 다양한 범위를 고려하여 위험 수준을 분류하고 고위험 수준의 인공지능 기술에 대해 윤리점검기준을 관리하는 등의 개입이 요구됨

4.3.3.2. 인공지능 윤리 준수 우수 기업체에 대한 장려책 제시

- 가) 인공지능 윤리에 관한 논의가 초기인 점을 고려하여 인공지능 자율점검표 활용이 우수한 기업체를 대상으로 각종 지원을 해 줌으로써 정부 시책을 적극 준수토록 유도할 수 있음

4.3.3.3. 인공지능 윤리에 대한 보편적 합의 유도

- 가) 윤리적 규범의 실천을 위해 인공지능 윤리 관리를 위한 거버넌스 구축과 실천을 촉진할 제도의 마련이 필요함
- 나) 인공지능 윤리에 관한 이슈에 대응하기 위한 국제적 규범화 노력에 공조할 필요가 있음.

4.3.4. 국내 주요 기업 사례 시사점

4.3.4.1. 2018년 이후 인공지능 윤리 관련 문헌 작성 발표

- 가) 국내 주요 기업들은 인공지능 윤리 현장, 원칙 또는 준칙, 규정, 점검표를 작성하고 발표하고 있음.
- 나) 대부분 기업들은 포괄적인 인공지능 윤리원칙을 제시하고 있으나 몇몇 기업들은 자사에서 개발 중인 특정 인공지능 서비스를 대상으로 윤리원칙을 발표함.

4.3.4.2. 정부의 인공지능 윤리원칙 부합

- 가) 기업의 윤리원칙은 대체로 인공지능 서비스를 개발하는 데 있어서 정부의 인공지능 윤리원칙의 핵심 요건에 부합되는 내용으로 발표함.
- 나) 인공지능 산업과 관련이 있는 국내 주요 기업들은 인공지능 윤리원칙을 발표하고 있지만, 이것의 실천에 관한 공시나 정보를 공개하지는 않음.

4.3.5. 소결론

4.3.5.1. 인공지능 윤리 기준 정립의 필요성 증대

- 가) 인공지능 정책 및 윤리 기준은 발표 시기에 차이가 있을 뿐 추구하고자 하는 방향은 동일함. 인간 중심적 사고의 대원칙 아래 구체적인 지침이나 하위 원칙으로 제시하고 있음
- 나) 국가차원이나 국제기구 차원에서 발표되는 인공지능 윤리체계 규범은 기업 차원에서 발표되는 규범보다 포괄적이며 더 넓은 범위를 다루고 있음. 다양한 이해관계자를 포용하고자 추상적인 단어와 내용위주로 구성되어 있음.

4.3.5.2. 국방 차원의 인공지능 윤리 기준 제정 시급

- 가) 민간 영역의 인공지능 윤리 기준을 준용하기보다는 국내 정부 인공지능 윤리 기준의 대전제 아래 국방분야에 특화된 기준을 정하는 것이 시급함.
- 나) 군 특성상 무기체계에 적용되는 인공지능의 기술과 자율무기체계에 적용할 수 있는 인공지능 윤리 기준을 정하는 것이 필요함.
- 다) 국방부는 정부기관으로서 국방분야에 적용할 인공지능 윤리 기준을 제시하고, 합참을 비롯한 각 군은 각 군의 특성을 반영한 윤리 기준의 이행 방안을 구체화하는 것이 바람직함.

4.3.5.3. 강제적 규범보다 자발적 참여 유도

- 가) 각 분야별 인공지능 추진 전략을 수립 시 인공지능 윤리에 대한 내용을 포함하고 정책 입안자와 사용자가 준수해야할 가이드 라인을 제시토록 유도.
- 나) 자율 점검표는 상황과 무기체계 특성에 맞추어 수정되고 변화될 수 있도록 유연성있게 작성.
- 다) 인공지능 관련 예산 분배시 국방 인공지능 윤리 기준에 부합되고 점검하는 무기 및 비무기체계에 대해서 예산 분배 우선권 부여 등 각 기관에서 자발적으로 참여할 수 있도록 제도적 정책 마련.

4.3.5.4. 국방 차원의 인공지능에 대한 공감대 형성 진행

- 가) 많은 국가에서 인공지능 윤리 기준을 제정시 각계의 의견과 토론을 거쳐 제정하였듯이 내부적인 토의와 외부 공청회 등을 통하여 한국군 특성에 맞는 윤리 기준 제정에 필요성을 공감하고 각 분야 전문가들의 의견을 수렴하는 과정이 필요함.

5. 국방 인공지능 윤리 기준(원칙) 제언

5.1. 윤리 기준 가이드 라인 설정

5.1.1. 국제인도법 (군사작전 3대 원칙)

5.1.1.1. 구별의 원칙

가) 전투원과 민간인의 구별, 군사시설 및 물자로부터 민간물자 및 시설을 구분하는 것이 중요함

5.1.1.2. 비례성의 원칙

가) 군사적 필요성에 의해 얻게 되는 군사적 이익과 공격으로 인해 발생하는 민간인의 부수적 피해를 비교하여 판단하는 것임. 이는 산술판단을 넘어선 외교적, 정치적 판단임

5.1.1.3. 사전주의 원칙

가) 언제나 인간이 개입할 수 있도록 자율무기체계를 설계하고 인간이 통제할 수 없는 상황을 최소화할 것을 요구함.

5.1.2. 전문가 및 담당자 의견

5.1.2.1. ‘데이터’의 윤리 기준은 ‘인공지능’윤리 기준에 포함하여 연구

- 가) 데이터의 윤리 기준의 필요성은 있으나 인공지능 윤리 기준의 하위 항목으로 관리
- 나) 별도의 범주로 관리하는 것은 중복소요가 많고 인공지능 윤리 기준과 데이터 윤리 기준을 각각 관리하기가 어려울 것으로 예상됨.

5.1.2.2. 국내 정부부처 인공지능 윤리원칙과 미 국방성 윤리 기준을 준용

- 가) 각 국과 해외 글로벌 기업 등의 윤리 기준을 살펴보면 유사한 항목들이 많음
- 나) 모든 항목을 고려하는 것보다 많이 준용되는 핵심 항목들 위주로 검토
- 다) 준용할 인공지능 윤리 기준으로 국내 정부부처 인공지능 10가지 원칙과 미 국방성 인공지능 기준 5가지를 참고하여 결정

5.1.2.3. 설문조사와 가중치 분석 기법을 통해 각 항목별 중요도 평가

- 가) 준용된 인공지능 윤리 기준 핵심 항목에 대해 설문조사로 우선 순위 판단
- 나) 우선순위 판단된 핵심항목에 대해 가중치 분석을 통해 계층적 구조 가시화

5.2. 윤리 기준 핵심 항목 도출

5.2.1. 핵심 항목 도출 절차

5.2.1.1. 기본 항목 도출

- 가) 국내 정부 기관에서 작성한 윤리 기준과 미 국방성의 윤리원칙, 해외 글로벌 기업의 인공지능 윤리 기준을 참고하여 주요 항목들을 정리함
- 나) 설문조사를 위한 적정 수량의 항목들로 요약하여 전문가 설문조사를 실시함. 국방분야에서 적용할 윤리 기준의 우선순위를 설정하고 국방 분야의 적정 인공지능 윤리 기준 항목 숫자도 병행하여 설문조사를 실시함
- 다) 의견 수렴된 항목들에 대해 가중치 분석을 통하여 평가항목별로 계층적 구조를 가시화함.
- 라) 이를 통해 국방분야에서 포함하여야 할 인공지능 윤리 기준 핵심 항목들을 도출해 낼 수 있음.

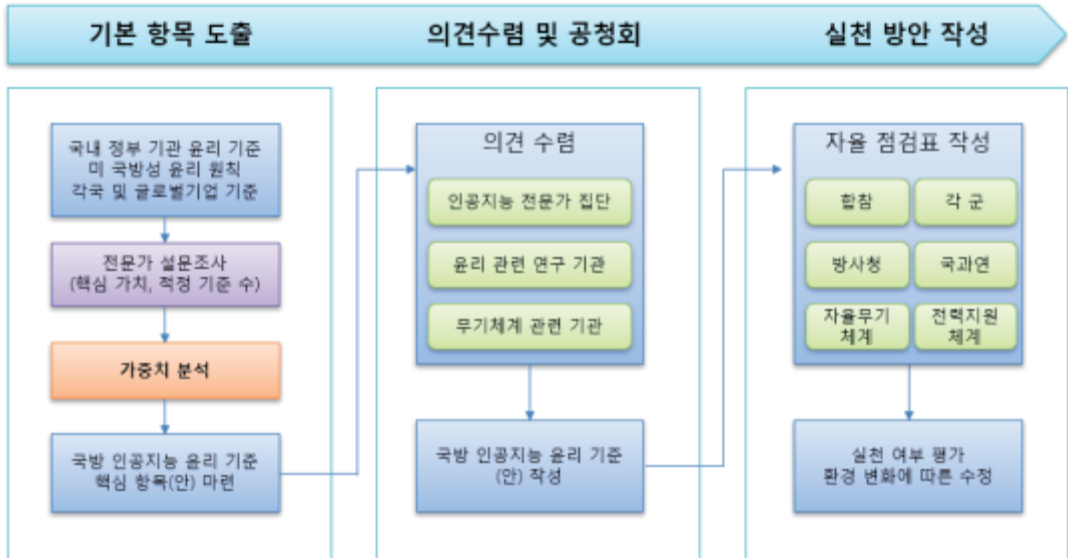
5.2.1.2. 의견 수렴 및 공청회

- 가) 관련 분야의 전문가들로 구성된 집단을 통해 국방분야 인공지능 윤리 기준 핵심 항목들에 대해 의견을 수렴하면서 필요성에 대해 공감대를 형성함
- 나) 인공지능 전문가 뿐만 아니라 윤리 관련하여 법학자들도 전문가 집단에 포함하며 무기체계 개발간 적용할 수 있도록 무기체계 관련 부서 근무자들도 의견 수렴 대상에 포함함
- 다) 이를 통해 국방 인공지능 윤리 기준(안)을 완성함

5.2.1.3. 실천 방안 작성

- 가) 제시된 국방분야 인공지능 윤리 기준을 실천하기 위한 점검표는 국방 인공지능 윤리 기준을 준수하기 위해 각 부대에서 제정하여 활용하도록 유도함.

- 나) 인공지능 관련 기술의 발전 추이를 고려하여 환경 변화에 따라 융통성있게 수정할 수 있도록 점검표 수정 권한을 각 부대로 위임하고 국방부는 조정·통제하는 역할을 수행함.



〈그림 5-1〉 국방 인공지능 기준 핵심 항목 도출 절차

5.2.2. 인공지능 윤리 기준 설문조사 항목 선정

5.2.2.1. 기존 사례별 인공지능 윤리 기준 종합

- 가) 아래의 도표는 연구보고서 IV장에서 조사한 인공지능 윤리 기준 범주별 분류 도표를 요약 정리한 것임
- 나) 예를 들어 주요 선진국 6개국의 인공지능 원칙 중 ‘책임성’이 6의 의미는 6개국 모두 채택하였다는 의미임.

구분	책임성	신뢰성 (안전성)	공평성 (공공성)	투명성 (추적, 설명)	데이터 관리	통제 가능성	인권보장 / 인간존중	프라이 버시 (개인정보)	기 타
주요 선진국 (6개)	6	5	5	6	2	2	2	4	16개 항목
국내 정부부처 (6개)	4	3	3	6	2	0	2	3	14개 항목
해외글로벌기업 (5개)	4	4	4	4	1	2	0	3	6개 항목
국내 주요기업 (5개)	2	3	3	4	1	0	2	3	10개 항목
계 (22개)	16	15	15	20	6	4	6	13	46개 항목

5.2.2.2. 윤리 기준 설문조사 대상 범위 선정

- 가) 사례 분석한 전체 약 60여개의 항목에 대해 설문조사하는 것은 비효율적임
- 나) 가장 빈도수가 높은 항목과 군사분야 유일한 사례인 미 국방성 사례(2020)를 주요 참고자료로 선정
- 다) 가장 빈도수가 높은 항목은 투명성(추적가능성, 설명가능성), 책임성, 신뢰성(안전성), 공평성(공공성), 프라이버시(개인정보) 순임
- 라) 또한 국방부는 정부기관으로서 국내 정부부처 합동 인공지능 윤리 기준(2020)을 참고할 필요가 있음을 고려
- 마) 국내 정부 부처 합동 인공지능 윤리 기준 안에 빈도수가 높은 대부분의 윤리원칙이 포함되어 있음

5.2.2.3. 설문조사 항목 선정

- 가) 국내 정부부처 인공지능 윤리 기준 원칙과 미 국방성 인공지능 원칙 중 중복된 항목(책임성) 제외 14개 항목으로 선정
- 나) 설문조사 대상 윤리 기준 : ① 인권존중, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성, ⑪ 공평성, ⑫ 추적가능성, ⑬ 신뢰성, ⑭ 통제가능성

5.2.3. 국내 전문가 및 담당자 대상 설문조사

5.2.3.1. 설문조사 개요

- 가) 국방분야 인공지능 원칙을 도출하기 위해 국내 전문가 및 인공지능 담당자를 대상으로 설문조사를 진행함.
- 나) 설문조사 기간은 2023. 8. 8 ~ 18일까지 11일간 진행함
- 다) 기간 중 총 참여 인원은 53명으로 국방부 / 국직부대 29명, 합참 / 각 군 19명, 학교 / 연구기관 5명이 참여하였음.

5.2.3.2. 설문조사 항목

- 가) 위에서 선정한 총 14개의 항목에 대해 우선순위 10개를 조사함.
- 나) 설문조사 대상자의 기본적인 인적 정보(소속기관, 소속부서, 직책)와 인공지능 윤리 기준 정립과 연관된 부서 근무 여부, 인공지능 연관 부서 근무 여부, 인공지능 윤리 기준 정립의 필요성, 인공지능 윤리원칙 우선순위별 10가지, 적절한 인공지능 윤리 기준 갯수 등을 조사함.

5.2.3.3. 설문조사 결과

- 가) 설문조사 응답 참여 인원은 총 53명으로 세부 현황은 아래 도표와 같음

〈표 5-1〉 인공지능 윤리 기준 설문조사 참여 인원

구분	국방부	국직기관	합참	육군	해군	해병대	학교기관	연구기관	계
명	16	13	2	12	4	1	3	2	53명
비율(%)	30%	24%	4%	23%	8%	2%	6%	4%	100%

- 나) 인공지능 윤리 기준 정립과 연관된 부서에서 근무하는 지 여부는 “그렇다”라고 응답한 응답자가 28명 (52.8%)이며, “그렇지 않다”라고 응답한 응답자는 25명 (47.2%)임. 인공지능 연관 부서에 근무하는 인원은 46명으로 전체의 86.8%로 대부분이 인공지능 연관부서에서 근무하는 인원이 설문조사에 참여하였음. 근무 경력은 약 80%가 5년 이내 근무자였음. 인공지능 윤리 기준의 필요성에 대해서는 모든 응답자가 인공지능 윤리 기준은 필요하다는 응답을 보였음. 이 중 절반 이상의 인원이 과거부터 인공지능 윤리 기준 필요성에 대해 인지하고 있었음을 알 수 있음.

다) 인공지능 윤리 기준 중 가장 우선시 되어야하는 항목으로 ‘통제가능성’을 선택하였고 그 다음으로 ‘신뢰성’, ‘책임성’, ‘안전성’, ‘추적가능성’, ‘인권보장’ 순으로 나타났다. 그 외 낮은 빈도로 ‘공공성’, ‘데이터관리’, ‘투명성’, ‘침해금지’ 등이 있음.

〈표 5-2〉 설문조사 결과 우선순위

구분	1순위	2순위	3순위	4순위	5순위	6순위	기타
항목	통제 가능성	신뢰성	책임성	안전성	추적 가능성	인권보장	공공성, 데이터 관리, 투명성, 침해금지 등
비율(%)	21.4	18.2	17.0	14.5	6.9	6.3	15.7

라) 인공지능 윤리 기준의 적절한 갯수는 4~5개가 38%, 6~8개가 34%로서 8개 이하가 적당하다는 의견이 전체의 77%를 차지함.

5.2.3.4. 설문조사 결과 분석

가) 설문조사 결과 윤리 기준으로서 포함해야 할 항목으로 ‘통제가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’ 등 4가지는 윤리 기준으로서 우선 순위가 높음을 알 수 있음.

나) 추가적으로 ‘추적가능성’, ‘인권보장’도 중요하게 생각하는 요소임을 알 수 있었음.

다) 윤리 기준 설정에 따른 키워드(갯 수)로는 6개가 적당하다는 의견이 대부분이었음.

라) 따라서 국방분야 인공지능 윤리 기준 우선 순위 설문조사 결과 윤리 기준 6개를 적용시 ‘통제가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’, ‘추적가능성’, ‘인권보장’ 등의 항목이 중요하다는 결과를 알 수 있었음.

마) 설문조사 제한사항으로 설문조사 항목에 대한 우선순위는 선정할 수 있었으나, 각 항목별 가중치 분석은 실시 할 수 없었음. 따라서 정확한 인공지능 윤리 기준 제정을 위해 핵심 키워드 가중치 분석¹⁰⁵⁾을 실시해야 함. 가중치 분석을 함으로써 인공지능 윤리 기준에 대한 핵심 키워드를 도출할 수 있음.

105) 전문가 집단을 대상으로 기준이 되는 두 개의 Factor간 쌍대비교 행렬화 연산을 통한 가중치 산정 기법

5.2.4. 우선순위 및 가중치 분석 결과

5.2.4.1. 가중치 분석(AHP¹⁰⁶⁾) 정의

- 가) 가중치 분석은 우선순위 도출을 위해 두 Factor간 쌍대비교 행렬화 연산을 이용한 가중치 산정 기반 의사결정 도구임.
- 나) 평가 항목들을 계층적으로 구성하고 쌍대비교를 통해 문제를 해결할 수 있는 기능을 제공함

5.2.4.2. 가중치 분석 목적

- 가) 국방 인공지능 실현을 위한 윤리적 측면의 원칙을 설정하여 핵심 키워드를 도출하고자 실시함.
- 나) ‘대한민국 인공지능 윤리 기준’을 기준으로 우선순위 및 가중치 분석, ‘미 국방성 인공지능 윤리원칙’을 기준으로 가중치 분석을 실시함.
- 다) 민간 측면 원칙 중 국방 측면에 반영되어야 하는 원칙 우선 순위 설정은 10점 척도를 적용함

5.2.4.3. 가중치 분석 설문 대상/실시

- 가) 인공지능 관련 부서 근무(업무) 유경험자를 대상으로 정책 부서, 각 군 관련 부서, 교육 및 연구기관 등 전문가 집단을 대상으로 실시함.
- 나) 국방부 데이터정책과 등 총 16명을 대상으로 설문 실시함.

5.2.4.4. 민간 측면에서 도출한 10개 핵심 키워드에 대한 우선순위 및 가중치 산출 결과

- 가) 핵심 키워드 우선순위 및 가중치 산출 결과 관계도는 <그림 5-2>와 같음.
- 나) AHP 분석에 있어서 CR 값이 0.1보다 작아야 설문에 일관성이 있으므로 즉, $CR < 0.1$ 인 설문을 이용하여 가중치의 평균을 산출(가중치는 기하평균)

106) AHP : Analytic Hierarchy Process

원칙(핵심 키워드)		순위	10점 척도 (점수비중)		항목	순위	가중치
민 간 측 면	인권보장	3	11.3%		인권보장	5	0.100
	프라이버시 보호	8	8.2%		프라이버시 보호	8	0.064
	침해금지	4	10.8%		침해금지	4	0.102
	다양성 존중	9	5.2%		다양성 존중	10	0.032
	공공성	6	10.5%		공공성	7	0.086
	연대성	9	5.2%		연대성	9	0.036
	데이터 관리	5	10.7%		데이터 관리	6	0.095
	책임성	2	13.2%		책임성	2	0.168
	안전성	1	15.5%		안전성	1	0.211
	투명성	7	9.3%		투명성	3	0.106
	총 계				100%	계	

〈그림 5-2〉 10점 척도와 AHP 가중치간의 관계도

- 다) 1, 2, 4순위는 동일하며, 3순위는 10점 척도에서는 ‘인권보장’, 가중치 분석에서는 ‘투명성’으로 분석됨
- 라) 우선순위에서 3순위로 선정된 ‘인권보장’은 가중치 분석에서는 5순위로 평가됨
- 마) 가중치 분석에서 3순위로 선정된 ‘투명성’은 10점 척도에서는 7순위로 낮게 평가됨
- 바) 도표의 적색 음영표시는 적용 항목의 경계선을 나타냄
- 사) 민간 측면에서 적용 중인 인공지능 윤리원칙 중 국방 인공지능 윤리원칙에 적용이 가능한 항목은 ‘책임성’, ‘안전성’, ‘침해금지’, ‘투명성’, ‘인권보장’, ‘데이터 관리’로 분석됨.

5.2.5. 미 국방성 인공지능 윤리원칙 항목에 대한 가중치 산출 결과

〈표 5-3〉 미 국방성 인공지능 윤리원칙 가중치 분석 결과

항목	가중치	내용
책임성	0.235	적절한 수준의 주의와 판단력을 지속적으로 행사(개발, 배치, 사용에 대한 책임도 유지)
공평	0.056	AI 성능에 있어 비의도적인 편향이 최소화 되도록 신중하게 조치
추적 가능성	0.112	관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적방법을 적절하게 이해할 수 있도록 개발
신뢰성	0.211	인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제 가능성	0.385	의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

※ AHP 분석에 있어서 CR 값이 0.1보다 작아야 설문에 일관성이 있으므로 즉, CR<0.1인 설문을 이용하여 가중치의 평균을 산출(가중치는 기하평균)

✓ 가중치 분석결과『통제가능성』>『책임성』>『신뢰성』>『추적가능성』>『공평성』순으로 중요도가 분석되었으며,『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단됨

5.3. 인공지능 윤리 기준(안) 제언

5.3.1. 국방 인공지능 윤리 기준(안) 도출

5.3.1.1. 미 국방성 인공지능 윤리원칙 가중치 분석(AHP) 결과

가) 『통제가능성』>『책임성』>『신뢰성』>『추적가능성』>『공평성』 순으로 중요도가 분석 되었으며, 『공평성』은 신뢰성에 포함될 수 있을 것으로 판단됨.

나) 즉, 『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단되어 추후 전문가 의견수렴 및 『신뢰성』의 정의 설정시 현 미 국방성 인공지능 윤리원칙 『공평성』에 정의되어 있는 편향성 부분을 반영하는 방안 등 토론(포럼, 세미나 등)을 통해 선정

5.3.1.2. 민간 측면에서 도출된 항목 중 미 국방성 인공지능 윤리원칙 적용 가능 항목

- 가) 우선순위 및 가중치 분석결과 1·2·4 순위로 동일하게 분석된 핵심 키워드 『안전성, 책임성, 침해금지』는 군사용 인공지능 윤리 기준에 반영 건의
- 나) 민간 측면에서 우선순위 5순위, 가중치 분석에서 6순위로 평가된 『데이터 관리』는 인공지능 학습의 기본이 되며 투명성 있는 데이터의 수집과 활용을 위해 군사용 인공지능 윤리 기준에 반영 건의
- 다) 민간 측면에서 우선순위 3순위, 가중치 분석에서 5순위로 평가된 『인권보장』 핵심 키워드는 군사적 활용 측면에서 적용 여부는 추가 검토가 필요(포럼, 세미나 등)
- 라) 『안전성』 → 미 국방성 인공지능 윤리원칙의 『신뢰성』에 포괄적으로 적용 가능
- 마) 『침해금지, 인권보장』 → 미 국방성 인공지능 윤리원칙의 『통제 가능성』에 포괄적으로 적용 가능
- 바) 『투명성』 → 미 국방성 인공지능 윤리원칙의 『추적 가능성』에 포괄적으로 적용

5.3.2. 국방 인공지능 윤리 기준 제언

- ✓ 통제 가능성, 책임성, 신뢰성(안전성, 공정성), 추적가능성(투명성), 데이터 관리, 인권보장

5.3.2.1. 통제가능성

- 가) 군사용 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생 시 사용자가 그 작동을 제어 또는 정지할 수 있는 기능을 갖추어야 한다.
- 나) 이상 징후 발생시 사용자가 쉽게 인지할 수 있어야 하며 이상 징후에 대한 경고 수준 설정을 할 수 있는 기능을 갖추어야 한다.
- 다) 인간의 의사 결정에 따라서 군사용 인공지능 시스템을 통제할 수 있어야 한다.
- 라) 인간의 통제를 벗어날 수 있는 잠재적 위험 요소 식별을 위해 노력하여야 한다.

5.3.2.2. 책임성

- 가) 군사용 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임 소재를 명확히 할 수 있어야 한다.
- 나) 군사용 인공지능의 사용, 개발, 배포, 결과에 대한 책임을 정의할 수 있어야 한다.

- 다) 인공지능 학습을 위해 수집·저장된 데이터에 대한 관리 책임을 정의할 수 있어야 한다.

5.3.2.3. 신뢰성(안전성, 공정성)

- 가) 외부 공격에 대한 회복성 등 시스템의 안정성을 보장해야 한다. 즉, 군사용 인공지능 시스템의 총수명주기간 성능의 보장, 안전, 보안 등 규정된 범위 내에서 시험되고 보증 되어야 한다.
- 나) 군사용 인공지능의 알고리즘에 대한 공정성 보장을 통해 공정한 의사결정(편향성 최소화)을 유도할 수 있어야 한다.

5.3.2.4. 추적 가능성(투명성)

- 가) 군사용 인공지능 알고리즘에 대한 의사결정 과정의 투명성 보장해야 한다. 즉, 군사용 인공지능 시스템은 의사결정 과정을 설명할 수 있도록 보장해야 한다.
- 나) 군사용 인공지능 시스템이 실패, 오작동 일으킬 경우에는 그 이유를 확인할 수 있어야 한다.
- 다) 군사용 인공지능은 설계로부터 전력화되는 모든 개발 과정을 이해관계자와 사용자에게 충분한 정보를 제공할 수 있어야 한다.

5.3.2.5. 데이터 관리

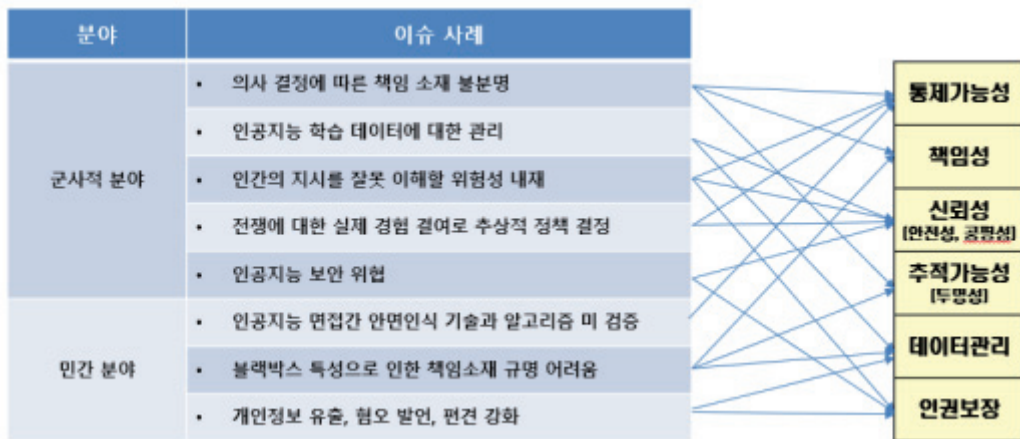
- 가) 군사용 인공지능 개발 및 활용 간 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- 나) 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발해야 한다.
- 다) 군사용 인공지능의 편향성(과대·과소 적합) 방지를 위한 데이터 품질 및 위험 관리 프로세스를 마련해야 한다.

45.3.2.6. 인권보장

- 가) 국제인도법을 준수하며 국제법에 명시된 권리를 보장해야 한다.
- 나) 인간을 수단으로 취급되지 않아야 하며, 기본적 권리와 자유를 침해하지 말아야 한다.

→ 제언한 군사용 인공지능 무기체계의 윤리 기준 적용시 <그림 5-3>과 같이 부작용 사례를 예방할 수 있을 것으로 판단된다.

✓ 부작용 사례와 국방부 인공지능 윤리 기준(안)의 관계



<그림 5-3> 국방인공지능 윤리 기준과 부작용 사례 관계

5.4. 인공지능 윤리 기준 평가방안

5.4.1. 개요

5.4.1.1. 윤리 기준 평가 목적

- 가) 국방부 인공지능 윤리 기준에 대해 국방분야 인공지능 분야 종사자가 고려해야 하는 요소와 이를 점검할 수 있는 방법을 제공함
- 나) 인공지능 윤리 기준이 현장에서 실현되고 각 담당자들의 자발적인 노력을 지원하기 위함.
- 다) 평가는 자율적으로 점검하되 각 분야별로 목적과 특성에 맞추어 수정이 가능함.

5.4.1.2. 평가 방법 종류

- 가) 규범적 정책을 이용하여 강제적으로 이행토록 제도화함

- 나) 권고사항으로 제시하며 각 기관별 점검 결과를 주기적으로 보고토록 함
- 다) 여기서는 자율적으로 점검할 수 있는 점검항목 위주로 제시함

5.4.1.3. 기존 인공지능 윤리 점검 기준표 참고

- 가) 육군본부 정책실에서는 ‘군사용 인공지능의 안전성 확보를 위한 단계별 윤리적 평가 및 검증요소(기준) 마련 연구’¹⁰⁷⁾를 실시하였음.
- 나) 과학기술정보통신부와 정보통신정책연구원은 ‘2023 인공지능 윤리 기준 실천을 위한 자율점검표(안)’¹⁰⁸⁾을 발표하였음.

5.4.2. 국방부 인공지능 윤리 기준별 점검표(안)

5.4.2.1. 점검 목적

- 가) 국방 인공지능 윤리 기준 자율점검표는 국방분야에서 개발되고 활용되는 인공지능 시스템의 국방 인공지능 윤리 기준을 준수하기 위한 핵심 항목을 실천하기 위함임
- 나) 국방 분야 인공지능 기술 발전을 저해하지 않으면서 전쟁법과 인간 중심의 인공지능 구현의 대원칙을 위배하지 않도록 예방하는 목적임.

107) 건양대학교 산학협력단, ‘군사용 인공지능의 안전성 확보를 위한 단계별 윤리적 평가 및 검증요소(기준) 마련 연구’, 2022.7.25

108) ‘과학기술정보통신부·정보통신정책연구원 「2023 인공지능 윤리 기준 실천을 위한 자율점검표 (안)」’

5.4.2.2. 점검표 구성

가) 점검표 앞서 제시한 핵심 키워드 6개에 대한 점검 항목으로 도출하였으며, 총 27개 항목으로 구성되어 있다.

〈표 5-4〉 국방 인공지능 윤리 기준 점검표 항목 수

핵심항목	통제가능성	신뢰성 (안전성, 공정성)	책임성	추적가능성 (투명성)	데이터관리	인권보장
코드번호	ME-01 ¹⁰⁹⁾	ME-02	ME-03	ME-04	ME-05	ME-06
문항 수	5	7(4+3)	4	5(2+3)	4	2

나) 통제가능성은 5가지 항목으로 시스템 측면, 관리 가능성, 접근 가능성, 일련의 절차, 통제 불능 최소화 등으로 구성되어 있다.

ME-01. 통제가능성

ME-01-01	군사용 인공지능시스템이 인간의 생명과 안전에 관한 권리를 침해하지 않도록 관리하는 시스템이 있는가?	예 아니오 미해당 ☐ ☐ ☐
ME-01-02	군사용 인공지능시스템이 인간의 생명과 안전에 관한 권리를 침해한다고 판단될 경우 이를 관리할 수 있는가?	☐ ☐ ☐
ME-01-03	군사용 인공지능시스템이 판단한 자율적 행동이나 결정에 수정할 경우 적절한 권한을 가진 사용자에게 의해 접근되고 통제될 수 있는가?	☐ ☐ ☐
ME-01-04	군사용 인공지능시스템의 개발·운영 단계에서 시스템의 목적에 맞도록 운영되는지 확인할 수 있도록 다양한 의견을 청취·검토·평가·반영할 수 있는 일련의 절차를 마련하였는가?	☐ ☐ ☐
ME-01-05	군사용 인공지능 시스템을 활용함에 있어서 인간이 통제할 수 없는 상황이 발생하지 않게하거나 최소화할 수 있는가?	☐ ☐ ☐

109) ME : Military Ethics

다) 신뢰성은 4가지 항목으로 근거 해석 가능성, 매뉴얼 제공 여부, 정보 제공 여부, 인간 존엄성 최소화 여부 등으로 구성되어 있다.

라) 신뢰성의 요소 중 안전성은 3가지 항목으로 안전 계획 작성, 사전 평가 및 완화 조치 여부, 평가 및 피드백 절차 여부 등으로 구성되어 있다.

ME-02. 신뢰성(안전성, 공정성)

ME-02-01	군사용으로 연구 및 활용될 인공지능 시스템을 개발하는 과정에서 중요한 결정에 대해서 사용자가 그 근거를 해석 및 추적할 수 있는가?	예 아니오 미해당 ☐ ☐ ☐
ME-02-02	군사 인공지능 시스템을 목적에 맞게 사용하기 위해 이해관계자와 사용자에게 관련 정보를 매뉴얼 등으로 제공하고 있는가?	☐ ☐ ☐
ME-02-03	군사용 인공지능시스템이 수집하는 데이터와 시스템의 의사결정에 영향을 주는 요인 등에 대한 정보를 제공하는 절차를 마련하였는가?	☐ ☐ ☐
ME-02-04	개발된 인공지능 성능이 인간의 존엄성을 해할 수 있는 비의도적인 편향이 최소화도록 설계되고 구현되었는가?	☐ ☐ ☐
ME-02-05	군사 인공지능 시스템의 개발 및 운용 과정에서 발생할 수 있는 중대한 위험(시스템 오류, 해킹, 악성 프로그램 설치, 적대적 공개 등)에 대응하기 위한 안전계획 또는 안전 조치(비상중단기능, 이용자 고지, 시스템 폐쇄 등)를 마련하고, 이를 실행할 수 있는 방법과한계에 대해 충분히 숙지한 사용자에게 의해서 운영하고 있는가?	☐ ☐ ☐
ME-02-06	군사 인공지능 시스템과 인간의 상호작용에서 발생가능한 과의존이나 인간 감시 등을 사전에 평가하고 완화하기 위한 조치를 마련했는가?	☐ ☐ ☐
ME-02-07	군사 인공지능 시스템의 결과에 대한 안전성을 지속적으로 평가하기 위한 전문가 평가 및 사용자 피드백 절차를 마련하였는가?	☐ ☐ ☐

마) 책임성은 4가지 항목으로 담당자 지정 여부, 역량 향상 노력 여부, 책임 소재, 배상 및 보상 준비 방안 마련 여부 등으로 구성되어 있다.

ME-03. 책임성		
ME-03-01	군사 인공지능 시스템을 개발 운영하는 과정에서 설계 및 개발자, 서비스 제공자, 사용자 간의 윤리 기준 준수를 위해 담당자를 지정하는 등 방안을 마련했는가?	예 아니오 미해당 ☐ ☐ ☐
ME-03-02	군사 인공지능 시스템 개발자 및 부서는 시스템이 적용되는 분야의 적합성을 판단하고 위험성을 인지하는 역량과 결정을 해석하는 역량을 향상시키기 위해 노력하고 있는가?	☐ ☐ ☐
ME-03-03	군사 인공지능 시스템의 개발 운영에서 발생하는 손해 또는 손실의 책임 소재를 명확히하고 있는가?	☐ ☐ ☐
ME-03-04	군사 인공지능 시스템의 사용과정에서 발생한 부당한 피해에 대하여 합리적 배상과 보상을 제공하기 위한 준비방안을 마련했는가?	☐ ☐ ☐

바) 추적가능성의 요소는 5가지 항목으로 결정 검토 절차, 판단 근거 해석, 정보 제공 여부, 의사결정 영향 요인 정보 제공 절차, 알고리즘 버전 관리 등으로 구성되어 있다.

ME-04. 추적가능성(투명성)		
ME-04-01	인공지능시스템의 결정에 대해 검토·평가·분석 할 수 있는 일련의 절차를 갖추고 있는가?	예 아니오 미해당 ☐ ☐ ☐
ME-04-02	군사 인공지능 시스템을 개발하는 과정에서 중요한 결정에 대해서 사용자가 그 근거를 해석 및 추적할 수 있는가?	☐ ☐ ☐
ME-04-03	군사 인공지능 시스템을 목적에 맞게 사용하기 위해 이해관계자와 사용자에게 관련 정보를 매뉴얼 등으로 제공하고 있는가?	☐ ☐ ☐
ME-04-04	군사 인공지능시스템이 수집하는 데이터와 시스템의 의사결정에 영향을 주는 요인 등에 대한 정보를 제공하는 절차를 마련하였는가?	☐ ☐ ☐
ME-04-05	군사용 인공지능시스템의 성능 향상을 위해 알고리즘이 수정될 경우 추적이 가능하도록 버전별로 관리되고 있는가?	☐ ☐ ☐

사) 데이터 관리는 4가지 항목으로 데이터 저장 여부, 접근 가능성과 분석 가능성, 업무 감독 절차, 성능지표 활용 여부 등으로 구성되어 있다.

ME-05. 데이터 관리

ME-05-01	군사용 인공지능시스템의 결정에 대해 추후에 검토·평가·분석할 수 있도록 운영과정에서 발생하는 모든 데이터를 저장하고 관리하는가?	예 아니오 미해당 ☐ ☐ ☐
ME-05-02	군사용 인공지능시스템 운영간 발생하는 저장되는 데이터에 대해 적절한 권한을 가진 사용자에게 의해 접근이 가능하고 분석이 가능하도록 관리하는가?	☐ ☐ ☐
ME-05-03	군사 인공지능시스템의 개발·운영에 활용되는 데이터의 수집과 처리 업무의 감독을 위한 절차를 마련하였는가?	☐ ☐ ☐
ME-05-04	군사 인공지능 시스템이 목적에 부합하는 예측을 수행하고 있는지를 데이터 관리 측면에서 평가하기 위한 적절한 성능지표를 사용하고 있는가?	☐ ☐ ☐

아) 인권 보장은 2가지 항목으로 국제인도법 준수 여부, 기본적 권리와 자유 침해 여부 등으로 구성되어 있다.

ME-06. 인권보장

ME-06-01	국제인도법을 준수하며 국제법에 명시된 권리를 보장하는가?	예 아니오 미해당 ☐ ☐ ☐
ME-06-02	인간을 수단으로 취급하지 않으며, 기본적 권리와 자유를 침해하지 않는가?	☐ ☐ ☐

6. 결론

6.1. 일반적인 윤리와 군사적 측면에서의 윤리

6.1.1. 일반적 개념의 윤리

- 윤리적 원칙들은 상호 연관성이 있으며, 사회적 조화와 공정한 교류를 이끌어내기 위한 기초 형성 또는 사회적 동의를 위한 거버넌스 구축을 추진하고 있다.
- 이를 개인, 조직, 사회 수준에서 적용하여 삶의 다양한 영역에서 윤리적인 행동을 실천하는 것이 중요하다.
- 일반적인 윤리의 기준은 자유주의의 윤리적 개념과 공동체주의의 도덕적 개념을 반영하여 사회 구성원들의 합의에 의해 형성된다고 할 수 있다.
- 즉, 일반적인 윤리는 사회적으로 공유되고 인정된 규범과 가치를 기반으로 한 행동의 올바름과 옳음에 대한 지침을 나타내는 것이며, 이는 개인, 집단, 조직 또는 사회 전체의 행동 및 결정에 영향을 미치고 상호 존중, 정의, 성실, 책임, 배려, 정직 등의 원칙에 근거한다.

6.1.2. 군사적 측면에서의 윤리

군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.

6.1.2.1. 군사용 인공지능의 주요 특성

- 가) 군사용 인공지능의 주요 목적은 군사력 향상, 안전 보장, 경제적 군 운영, 민첩한 대응 및 전략적인 운용을 위해 기술적인 혁신과 응용을 통합하여 활용하는 것으로 이러한 기술의 개발과 사용에는 윤리적, 법적인 측면에서도 주의가 필요하다.
- 나) 군사용 인공지능 자율무기체계의 개념 및 범위는 군사 로봇에 포함된 자동화 무기(Autonomous Weapons)를 의미하며, 이를 통틀어 치명적인 자동화 무기체계(LAWs, Lethal Autonomous Weapons Systems) 혹은 자율 살상 무기라 부르기도 한다.
- 다) 군사용 인공지능 자율무기체계의 유형은 자동적 방어체계와 자동적 공격체계로 분류할 수 있으며, 인간이 배제된 신속한 대응을 할 수 있는 체계이다.

- 라) 자율무기체계의 윤리적 쟁점은 『위협의 크기, 특성, 양상의 변화 - 기존 위협의 확대, 위협의 특성 변화, 새로운 위협의 출현』과 『기계에 의한 의사결정 및 살인의 책임소재가 모호하다는 점』, 『전쟁에 대한 의사결정이 추상적이 될 위협』 측면에서 대두되고 있다.
- 마) 특히 자율살상무기(LAWs)의 윤리적 쟁점은 『법적·윤리적 측면』, 『윤리적·도덕적 해이의 문제』, 『생명권 경시의 문제』, 『인간의 존엄과 가치의 문제』, 『인간성의 피폐화와 인간의 수단화 문제』 등으로 대두되고 있다.
- 바) 국제법에 따른 자율무기체계에 대한 윤리적 쟁점은 『국제인도법(군사작전 3대 원칙-구별의 원칙, 비례성의 원칙, 사전주의 원칙)』, 『과잉무기사용금지 원칙』, 『무력충동 원칙 위배 책임』 등으로 대두되고 있다.

6.1.2.2. 인공지능 자율무기체계와 자율살상무기체계

- 가) 자율무기체계는 물리적 교전, 전자 공격, 사이버 방호 등의 다양한 임무에 활용되며 인간 운용자의 관제를 받는 자율체계까지 포함된다.
- 나) 자율살상무기체계의 구체적인 정의에 대해서는 국제적으로 합의된 바 없으나 자율살상무기는 자율성과 교전권이 부여된 특정 자율무기체계를 나타내는 포괄적인 개념으로 정의한다.

6.1.2.3. 군사적 측면에서 윤리적 논란

가) 자율살상무기체계 규제를 위한 국제적 논의

- (1) 자율살상무기체계의 개발과 배치, 윤리적 문제를 최초로 제기한 국제기구인 국제 로봇 무기 통제 위원회(ICRAC)로 자율무인전투체계의 연구·개발, 배치, 사용을 금지하는 의제에 대하여 국제적 논의를 제안
- (2) 비정부기구로서 살인 로봇 반대 캠페인(Campaign to Stop Killer Robots)과 국제적십자위원회(ICRC)에서 국제인도법에 위배될 수 있다는 문제점을 제기
- (3) 또한, 국제 인권 감시기구(Human Rights Watch)와 국제 앰네스티(Amnesty International)도 전장에서 활용할 때에 국제인도법을 준수하고 윤리적 행동을 위해 선제적 예방조치가 필요하다고 주장
- (4) 유엔은 2013년 유엔 인권 이사회(UN Human Rights Council)에 제출된 특별보고서에는 회원국들의 국제적 합의가 있을 때까지 자율살상무기체계의 개발, 생산, 배치를 중지할 것을 권고

- (5) 2014년 유엔의 CCW 회의에서 자율살상무기체계의 개념과 정의, 기술적 특성, 윤리적·안보적 이슈, 국제 규범화 등이 논의되었으나 이에 대한 구체적인 실행 계획이 도출되지 못함

나) 자율살상무기(LAWs) 개발·사용·생산에 대한 국제사회의 동향 및 찬·반론

- (1) 찬성론: 방위비 부담을 줄이고 인명 손실을 최소화하자는 의견, 윤리적 측면에서 꼭 필요한 명령만 내려 로봇 무기가 무고한 살상을 방지할 수 있는 프로그램을 개발하면 피해를 최소화할 수 있는 장점, LAWs에 의한 살인 책임 책임자의 경계는 누가 명령을 내렸는지 기록하기가 더 쉬워 책임자도 명확하게 가려낼 수 있다고 주장
- (2) 반대론: 킬러-로봇이 전투의 문턱을 낮춰 인간의 삶을 크게 훼손할 것이라는 부분, 전 세계적으로 영원한 전쟁 그리고 어떠한 법적 책임이나 민주적 논의도 없는 수많은 비밀 공작 전쟁이 될 것이라고 경고, 근본적인 법적 및 윤리적 문제(이성↔법), 국제법 위배 등으로 반대

6.1.2.4. 윤리적 논란 해소방안

- 기존 제안된 인공지능 윤리 권고안들은 추상적 원칙들로 이루어져 있어 구체적인 실행 방안이 필요하다.
- 미 국방부가 2020년 2월 발표한 인공지능에 대한 윤리원칙 및 적용하는 가이드 라인이 제시, 한국 군에서도 면밀한 검토를 통해 한국 군사 상황에 맞는 군사 인공지능 윤리 개발 가이드 라인이 필요하다.
- 실질적인 해소 방안으로 윤리적인 설계 및 검수를 거친 군사용 인공지능은 인간의 통제하에 있을 때 오히려 전시에서의 인명 훼손을 이전보다 감소시킬 수 있으므로, 인공지능 무기의 기획, 설계, 개발 등의 전체 단계를 통해 윤리적 점검사항들을 정하여 적용할 필요(인공지능 윤리 자율 점검 항목들로 구현될 필요가 있음)하다.
- 자율살상무기가 완전한 자율시스템의 수준이 될 때 피아식별이나 전투원-민간인 구분에 있어 인간보다 훌륭한 성능을 보일 수 있는지에 대한 문제이다.
- 가장 높은 단계의 자율성을 가지는 인공지능 수준에서 제기될 문제로 이에 앞서 인간의 제어가 필수적인 반(semi)자율 인공지능에서의 윤리적 설계와 인간의 역할을 구체화할 필요성이 있다.
- 현 한국 군사 관련 현실적 상황 맥락을 검토하여 한국의 이익을 고려하여 수정 및 재구성하는 것이 필요하다.
- 인공지능 편향은 반드시 부정적인 것이 아니며 군사용 무기라는 특정 목적을 가지는 자율무기체계에 있어서는 데이터 및 알고리즘 편향이 인공지능시스템 설계 과정에서 자국의 이익을 침해하는 부분과 이익에 도움이 되는 부분을 분리할 필요성이 있으며, 이는 군사적 특성마다 달라지므로 한국의 자국과 적국의 정의 및 그 특성에 따라 구성될 필요가 있다.

6.2. 인공지능 윤리 이론 및 사례

6.2.1. 인공지능 윤리 관련 이론

6.2.1.1. 인공지능에서 말하는 윤리

- 가) 인공지능 시스템이 가지고 있어야 하는 윤리적 원칙, 가치, 행동 지침을 의미
- 나) 인공지능은 점점 더 사회 시스템에서 중요한 역할을 수행하며 사회, 경제, 문화, 정치 등 다양한 영역에 영향을 미치게 되면서 그 사용과 개발에 대한 윤리적 고려가 더욱 중요해지고 있는 실정이다.

6.2.1.2. 인공지능의 윤리원칙

- 가) 공정성과 투명성, 개인정보 보호와 프라이버시, 안전과 보안, 책임과 책임 소재, 사회적 영향과 공익, 자율성과 의사결정 투명성, 사람 중심적 가치, 지속 가능성이다.
- 마) 이러한 윤리적 원칙은 인공지능 개발자, 이용자, 정부, 기업 및 기술 커뮤니티에게 인공지능의 윤리적인 사용과 발전을 촉진하며, 사회적으로 수용 가능한 방향으로 인공지능 기술을 유도해야 한다.

6.2.1.3. 인공지능 윤리의 필요성

- 가) 인공지능 기술의 급속한 발전으로 제조·의료·교통·환경·교육 등 산업 전 분야에 인공지능이 활용·확산되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 인공지능 윤리 이슈가 새롭게 대두되고 있다.
- 나) 인공지능에 대한 법적 규제의 불충분성으로 규범적 문제에 대응하는 법적 규제가 마련되지 않았다는 점이다.
- 다) 윤리를 통한 자율적 규제는 인공지능 과학기술이 원활하고 창의적으로 발전할 수 있도록 하면서도, 다른 한편으로 인공지능이 유발하는 규범적 문제에 적절하게 대처하기 위해서는 법적 규제보다 직접성이나 강도가 약한 윤리적 규제를 고려할 필요가 있다는 것으로 인공지능과 윤리 문제를 논의할 필요가 있다.

6.2.1.4. 인공지능 윤리에서 말하는 윤리 개념

- 가) 인공지능 윤리에서 말하는 윤리는 자유주의가 강조하는 도덕 개념으로 한정하여 이해하는 것이 적절하다고 판단된다.
- 나) 현 시점에서 구현되는 있는 인공지능은 약한 인공지능으로 인간에 대한 관계에서 옳음을 준수하는 것으로 인간의 권리나 이익을 침해하지 않는 것이라 할 수 있다.
- 다) 즉, 인공지능에게 요구해야 하는 윤리는 그 의미 내용이 포괄적이면서도 모호한 윤리보다는 외연이 좁으면서도 그 의미 내용이 구체적인 도덕으로 보아야 한다.

6.2.1.5. 윤리의 구별 개념

- 가) 인공지능 윤리는 공학윤리 가운데 인공지능을 대상으로 하여 전문적으로 특화된 영역이라고 할 수 있으나, 인공지능 윤리는 인공지능을 설계하는 설계자에게 적용되는 윤리뿐만 아니라 인공지능 자체가 준수해야 하는 윤리, 다시 말해 인공지능 자신을 위한 윤리까지 논의 대상에 포함된다.
- 나) 로봇 윤리와 인공지능 윤리를 비교하면 인공지능은 인간의 의식, 즉 의식체계를 인공적으로 구현한 것에 해당되며, 로봇은 인간의 몸 즉, 생명체계를 인공적으로 구현한 것으로 구별되지만, 인공지능과 로봇이 결합되어 등장하는 오늘날의 상황을 고려할 때 양자를 거의 같은 의미로 사용하는 것도 문제가 없을 것으로 판단된다.

6.2.1.6. 인공지능 윤리의 수범 주체 및 기본방향

- 가) 인공지능 윤리 제정시 수범 주체는 현시점에서 인공지능 개발자 위주의 제정이 우선시 되어야 하며, 향후 인공지능 발전에 따라서 인공지능 자체 및 이용자에 대한 윤리 기준 및 법적인 인정 부분도 고려해야 한다.
- 나) 결과적으로 인공지능 윤리는 정확하게 말하면 인공지능 개발자가 준수해야 하는 윤리, 인공지능 개발자의 윤리라고 보는 것이 타당하다.

다) 기본방향

- (1) 인공지능 윤리는 자유주의 윤리를 지향해야 하며, 이는 자유주의 윤리의 규범적 의미가 가장 구체적이고 명확하기 때문이다.
- (2) 자유주의 윤리를 가장 우선시 하면서도 보완적으로 공리주의 윤리를 수용할 필요가 있으며, 이는 사람과 사람 사이의 자유나 권리가 충돌하는 상황에서 서로의 이익을 최적화하기 위해서는 이익형량적 사고를 어느 정도 수용해야 함.

- (3) 기본 방향 설정에 따른 윤리원칙: 인간의 존엄성 존중, 권리 침해 금지, 실정법 위반 금지, 이용자의 명령 복종, 투명성, 설명 가능성, 책임

6.2.2. 인공지능 부작용 사례(윤리적 측면)

6.2.2.1. 인공지능 투명성, 인공지능 공정성, 인공지능 오작동, 인공지능 보안위협, 기타 부작용 사례 등이 있음.

- ✓ 자율무기체계의 사이버 공격 대응 부재 및 부족, 전쟁 발발 가능성 증대, 자율무기체계의 목적 외 용도로 사용, 생명 경시 및 인간 존엄 훼손 등 이슈 발생 우려

6.2.2.2. 군사용 인공지능 자율무기체계의 과대적합/과소적합에 따른 문제점

가) 과대 적합·과소 적합은, 학습 반복 수준과 학습 데이터의 분포에 따라 그 결과가 다르게 나타나므로 국방 분야의 특징상, 일부 분야(ex. 전장, 전투 방식 등)에 대한 학습데이터를 확보하기 어렵기 때문에 인공지능 적용 분야를 고려할 필요가 있음.

나) 이러한 현상은 무기체계의 상태 진단 오류, 부품 교체 주기 부적절, 자율무기체계의 미작동 및 오작동, 피아식별 오류 등 발생할 수 있음.

6.3. 기 제정된 인공지능 윤리 기준(원칙) 분석을 통한 시사점

6.3.1. 인공지능 윤리 정책 동향 분석 및 주요 선진국 사례

6.3.1.1. 인공지능 윤리 정책 동향

가) 세계 각국과 글로벌 기업 및 주요 국제기구들은 인공지능의 활용도 증가에 따른 산업의 진흥을 모색하는 한편 윤리적인 인공지능 실현을 위한 정책을 수립해 오고 있다.

나) 인공지능 윤리에 대한 필요성을 인식하고 해외 글로벌 기업을 선두로 인공지능 기술에 대한 원칙을 제정하기 시작한 이후 인공지능 학회를 중심으로 각종 이슈를 정리하고 윤리 및 가치를 논의하기 시작했으며 장기적인 관점에서의 원칙들을 발표하기 시작하였다.

- 다) 인공지능 기술과 인공지능 기반 제품 및 서비스가 활성화되고 윤리적 이슈와 개인정보 유출과 같은 문제가 지속 제기됨에 따라 인공지능 선도 국가들은 국가 정책으로 인공지능 활용을 위한 정책을 마련하여 발표하였다.
- 라) 최근 우리나라도 인공지능 윤리 기준 및 원칙을 발표하고 기존에 발표한 인공지능 윤리 기준이 선언적으로만 그치지 않도록 각종 가이드 라인, 개발 기준 등을 마련하고 있다.
- 마) 군사 관련 분야에서는 미국 국방성에서 국제인도법의 원칙을 준수하는 범위 내에서 군사 분야의 특성을 고려하여 국방분야에 적용할 인공지능 윤리원칙을 제시하였다.

6.3.1.2. 주요 선진국 사례

- 가) 아실로마 AI 원칙은 인공지능 개발과 관련하여 안전, 투명성, 책임, 인간의 가치, 개인정보 보호, 공동 이익 등의 윤리적 기준을 제시하고 있다.
- 나) EU, ALTAI 발표 (20. 7.)
 - (1) 총 143개의 방대한 체크리스트 문항을 포함하여, 산업별·기업별 다른 특성을 가진 기업들이 자신의 상황에 맞게 활용할 수 있도록 작성
 - (2) 자율성을 최대화하고 인공지능이 미칠 수 있는 잠재적인 영향에 대한 인식을 제고
- 다) 영국, 데이터 윤리 프레임워크 발표 (20. 9.)
 - (1) 원칙의 적용과 행동 실천을 구체적 점수로 평가하고, 원칙 중 하나라도 3점 이하의 평가를 받을 경우 문제가 있음을 명시하며, 이에 대한 팀내 또는 관련 그룹과의 논의를 통해 해결책을 찾아야 함을 강조
 - (2) 실질적으로 실무자들이 확인할 수 있는 수치적 근거를 제시할 수 있도록 하였음
- 라) OECD, 인공지능 신뢰성 확보를 위한 도구 발표 (21. 6.)
 - (1) 개별 도구의 품질이나 완성도는 평가하지 않았으며, 정책 입안자 또는 실무자가 필요한 도구를 신속하게 식별할 수 있도록 참고자료를 제공하는 데 중점을 두고 제작
 - (2) 인공지능 도구를 유형별로 분류·분석한 후, 각 도구들의 정보를 구조화하고, 목적이나 활용 맥락이 서로 다른 도구들을 비교하기 위한 프레임워크 개발
- 마) UNESCO, 인공지능 윤리 권고안 발표 (21. 11.)
 - (1) UNESCO 회원국과 권고안에 언급된 모든 이해관계자들이 해당 권고안에 명시

된 윤리적 가치, 원칙을 존중·증진·보호하고 권고 정책의 시행에 필요한 실현 가능한 모든 조치를 취해야 함을 강조

바) 일본, 인공지능 원칙 실천을 위한 거버넌스 가이드 라인 발표 (‘22. 1.)

- (1) 일본 경제산업성은 본 가이드 라인이 살아있는 문서라는 점에 방점을 두고 있음.
- (2) 인공지능 기술 발전이 여전히 활발하게 진행되고 있다는 점을 반영하면서, 인공지능의 개발·활용을 사회적으로 수용 가능한 수준에서 관리하기 위함임

사) 미국, 인공지능 권리장전 청사진 발표 (‘22. 10.)

- (1) 인공지능 권리장전 청사진은 프레임워크로서 기존 미국 내에서 제시되고 있는 인공지능 사용에 관련된 원칙을 전반적으로 정리한 것임.
- (2) 단순히 원칙만을 제시한 것이 아니라 해당 원칙이 현실에서 실천으로 어떻게 이행될 수 있는지, 미국 연방정부, 국립연구소 등에서 실시하고 있는 각종 연구나 리스크 관리 프레임워크, 관련 표준에 대한 지침과 더불어 민간 사례에 대한 분석을 제공한다는 점에서 의미가 있음.

아) 미국 국방성, 인공지능 윤리원칙 발표 (‘20. 2.)

기본 원칙	주요 내용
책임성	- 적절한 수준의 주의와 판단력을 지속적으로 행사 (개발, 배치, 사용에 대한 책임도 유지)
공평성	인공지능성능에 있어 비의도적인 편향이 최소화되도록 신중하게 조치
추적가능성	관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발
신뢰성	인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제가능성	의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

6.3.2. 국내·외 사례 분석을 통한 시사점

6.3.2.1. 해외 글로벌 기업 사례 시사점

가) 인공지능 기술 발전에 따른 부작용 우려

- (1) 인공지능 기술과 서비스가 일상화되고 보편화되는 가운데 잠복해 있던 윤리 훼손 이슈에 대해 관심도가 높아졌음.

- (2) 인공지능을 더욱 적극적으로 개발하고 활용범위를 최대한 확대하고자 할수록 인공지능 활용을 통한 이익의 확보와 인공지능에 대한 국민의 보호를 균형적으로 이루어야 함.

나) 인공지능 윤리의 확산

- (1) 인공지능 서비스로 인한 부작용이나 위험을 최소화하기 위해서는 기업의 사회적 책임에 수반되는 윤리의식의 선행이 필요하다는 공감대가 형성됨
- (2) 인공지능의 결함을 해결하지 못하고 인공지능 서비스를 시장에 출시할 경우 엄청난 법적, 재무적 위험에 직면한다는 현실을 직시하게 되었음.
- (3) 이러한 법적, 재무적 손실과 더불어 기업은 기존에 쌓아왔던 명성과 같은 유·무형 손실이 가중될 수 있다는 것을 인지함.

6.3.2.2. 국내 정부 부처 사례 시사점

가) 인공지능 윤리 기준 제정을 위한 정부의 적절한 개입 필요

- (1) 인공지능 윤리원칙 준수를 위한 기업의 자율적 노력도 필요하지만 관련 정보를 교환하고 자율적 규정을 마련할 수 있도록 정부의 가이드 라인 제정이 필요함
- (2) 인공지능 기술이 적용되는 다양한 범위를 고려하여 위험 수준을 분류하고 고위험 수준의 인공지능 기술에 대해 윤리점검기준을 관리하는 등의 개입이 요구됨

나) 인공지능 윤리 준수 우수 기업체에 대한 장려책 제시

- (1) 인공지능 윤리에 관한 논의가 초기인 점을 고려하여 인공지능 자율점검표 활용이 우수한 기업체를 대상으로 각종 지원을 해 줌으로써 정부 시책을 적극 준수토록 유도할 수 있음

다) 인공지능 윤리에 대한 보편적 합의 유도

- (1) 윤리적 규범의 실천을 위해 인공지능 윤리 관리를 위한 거버넌스 구축과 실천을 촉진할 제도의 마련이 필요함
- (2) 인공지능 윤리에 관한 이슈에 대응하기 위한 국제적 규범화 노력에 공조할 필요가 있음.

6.3.2.3. 국내 주요 기업 사례 시사점

가) 2018년 이후 인공지능 윤리 관련 문헌 작성 발표

- (1) 국내 주요 기업들은 인공지능 윤리 헌장, 원칙 또는 준칙, 규정, 점검표를 작성하고 발표하고 있음.

- (2) 대부분 기업들은 포괄적인 인공지능 윤리원칙을 제시하고 있으나 몇몇 기업들은 자사에서 개발 중인 특정 인공지능 서비스를 대상으로 윤리원칙을 발표함.

나) 정부의 인공지능 윤리원칙 부합

- (1) 기업의 윤리원칙은 대체로 인공지능 서비스를 개발하는 데 있어서 정부의 인공지능 윤리원칙의 핵심 요건에 부합되는 내용으로 발표함.
- (2) 인공지능 산업과 관련이 있는 국내 주요 기업들은 인공지능 윤리원칙을 발표하고 있지만, 이것의 실천에 관한 공시나 정보를 공개하지는 않음.

6.3.3. 종합 시사점

6.3.3.1. 인공지능 윤리 기준 정립의 필요성 증대

- 가) 인공지능 정책 및 윤리 기준은 발표 시기에 차이가 있을 뿐 추구하고자 하는 방향은 동일함. 인간 중심적 사고의 대원칙 아래 구체적인 지침이나 하위 원칙으로 제시하고 있음
- 나) 국가차원이나 국제기구 차원에서 발표되는 인공지능 윤리체계 규범은 기업 차원에서 발표되는 규범보다 포괄적이며 더 넓은 범위를 다루고 있음. 다양한 이해관계자를 포용하고자 추상적인 단어와 내용위주로 구성되어 있음.

6.3.3.2. 국방 차원의 인공지능 윤리 기준 제정 시급

- 가) 민간 영역의 인공지능 윤리 기준을 준용하기보다는 국내 정부 인공지능 윤리 기준의 대전제 아래 국방분야에 특화된 기준을 정하는 것이 시급함.
- 나) 군 특성상 무기체계에 적용되는 인공지능의 기술과 자율무기체계에 적용할 수 있는 인공지능 윤리 기준을 정하는 것이 필요함.
- 다) 국방부는 정부기관으로서 국방분야에 적용할 인공지능 윤리 기준을 제시하고, 합참을 비롯한 각 군은 각 군의 특성을 반영한 윤리 기준의 이행 방안을 구체화하는 것이 바람직함.

6.3.3.3. 강제적 규범보다 자발적 참여 유도

- 가) 각 분야별 인공지능 추진 전략을 수립 시 인공지능 윤리에 대한 내용을 포함하고 정책 입안자와 사용자가 준수해야 할 가이드 라인을 제시토록 유도.
- 나) 자율 점검표는 상황과 무기체계 특성에 맞추어 수정되고 변화될 수 있도록 유연성있게 작성.

- 다) 인공지능 관련 예산 분배시 국방 인공지능 윤리 기준에 부합되고 점점하는 무기 및 비무기체계에 대해서 예산 분배 우선권 부여 등 각 기관에서 자발적으로 참여할 수 있도록 제도적 정책 마련.

6.3.3.4. 국방 차원의 인공지능에 대한 공감대 형성 진행

- 가) 많은 국가에서 인공지능 윤리 기준을 제정시 각계의 의견과 토론을 거쳐 제정하였듯이 내부적인 토의와 외부 공청회 등을 통하여 한국군 특성에 맞는 윤리 기준 제정에 필요성을 공감하고 각 분야 전문가들의 의견을 수렴하는 과정이 필요함.

6.4. 국방 인공지능 윤리 기준(원칙) 제언

6.4.1. 윤리 기준 가이드 라인 설정 및 핵심 항목 도출을 위한 절차

6.4.1.1. ‘데이터’의 윤리 기준은 ‘인공지능’ 윤리 기준에 포함하여 연구

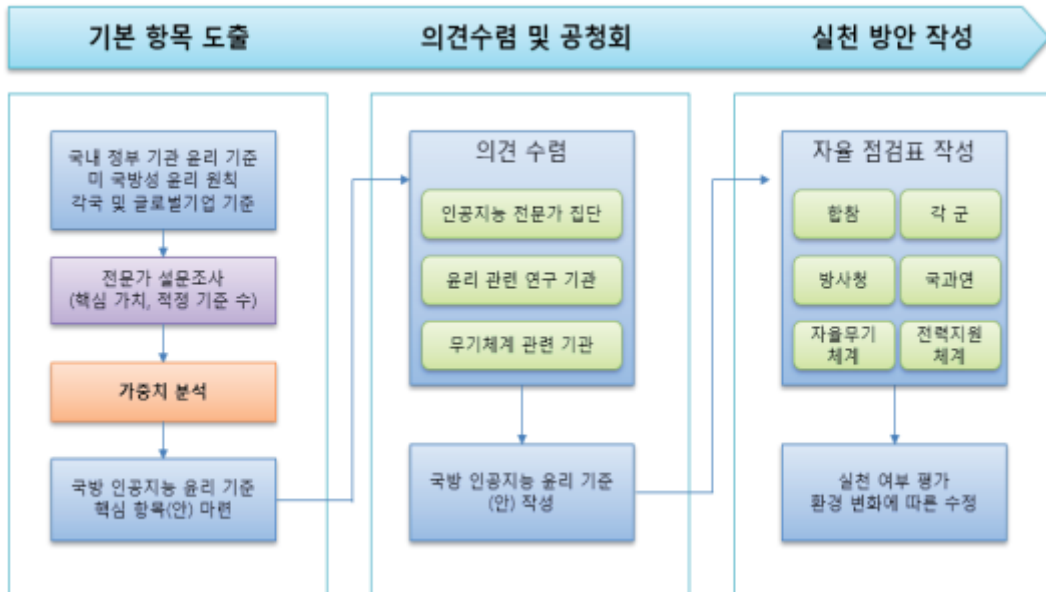
- 가) 데이터의 윤리 기준의 필요성은 있으나 인공지능 윤리 기준의 하위 항목으로 관리
- 나) 별도의 범주로 관리하는 것은 중복소요가 많고 인공지능 윤리 기준과 데이터 윤리 기준을 각각 관리하기가 어려울 것으로 예상됨.

6.4.1.2. 국내 정부부처 인공지능 윤리원칙과 미 국방성 윤리 기준을 준용

- 가) 각 국과 해외 글로벌 기업 등의 윤리 기준을 살펴보면 유사한 항목들이 많음
- 나) 모든 항목을 고려하는 것보다 많이 준용되는 핵심 항목들 위주로 검토
- 다) 준용할 인공지능 윤리 기준으로 국내 정부부처 인공지능 10가지 원칙과 미 국방성 인공지능 기준 5가지를 참고하여 결정

6.4.1.3. 설문조사와 가중치 분석 기법을 통해 각 항목별 중요도 평가

- 가) 준용된 인공지능 윤리 기준 핵심 항목에 대해 설문조사로 우선 순위 판단
- 나) 우선순위 판단된 핵심항목에 대해 가중치 분석을 통해 계층적 구조 가시화



6.4.2. 설문조사 및 가중치 산출 결과

6.4.2.1. 국방 인공지능 윤리 기준(안) 설정을 위한 항목별 설문조사 결과

- 가) 설문조사 결과 윤리 기준으로서 포함해야 할 항목으로 ‘통제가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’ 등 4가지는 윤리 기준으로서 우선 순위가 높음을 알 수 있음.
- 나) 추가적으로 ‘추적가능성’, ‘인권보장’도 중요하게 생각하는 요소임을 알 수 있었음.
- 다) 윤리 기준 설정에 따른 키워드(갯 수)로는 6개가 적당하다는 의견이 대부분이었음.
- 라) 따라서 국방분야 인공지능 윤리 기준 우선 순위 설문조사 결과 윤리 기준 6개를 적용시 ‘통제가능성’, ‘신뢰성’, ‘책임성’, ‘안전성’, ‘추적가능성’, ‘인권보장’ 등의 항목이 중요하다는 결과를 알 수 있었음.

구분	1순위	2순위	3순위	4순위	5순위	6순위	기타
항목	통제 가능성	신뢰성	책임성	안전성	추적 가능성	인권보장	공공성, 데이터 관리, 투명성, 침해금지 등
비율(%)	21.4	18.2	17.0	14.5	6.9	6.3	15.7

마) 설문조사 제한사항으로 설문조사 항목에 대한 우선순위는 선정할 수 있었으나, 각 항목별 가중치 분석은 실시 할 수 없었으며, 이에 정확한 인공지능 윤리 기준 제정을 위해 핵심 키워드 가중치 분석을 실시함.

6.4.2.2. 핵심 키워드에 대한 가중치 분석결과

가) 가중치 분석은 우선순위 도출을 위해 두 Factor간 쌍대비교 행렬화 연산을 이용한 가중치 산정 기반 의사결정 도구로 평가 항목들을 계층적으로 구성하고 쌍대비교를 통해 문제를 해결할 수 있는 기능을 제공함(AHP 기법 활용)

나) 민간 측면에서 도출한 10개 핵심 키워드에 대한 우선순위 및 가중치 산출 결과 아래 그림과 같음.

원칙(핵심 키워드)		순위	10점 척도 (점수비중)		항목	순위	가중치
민 간 측 면	인권보장	3	11.3%		인권보장	5	0.100
	프라이버시 보호	8	8.2%		프라이버시 보호	8	0.064
	침해금지	4	10.8%		침해금지	4	0.102
	다양성 존중	9	5.2%		다양성 존중	10	0.032
	공공성	6	10.5%		공공성	7	0.086
	연대성	9	5.2%		연대성	9	0.036
	데이터 관리	5	10.7%		데이터 관리	6	0.095
	책임성	2	13.2%		책임성	2	0.168
	안전성	1	15.5%		안전성	1	0.211
	투명성	7	9.3%		투명성	3	0.106
총 계			100%		계		1.000

- (1) 1, 2, 4순위는 동일하며, 3순위는 10점 척도에서는 '인권보장', 가중치 분석에서는 '투명성'으로 분석됨.
- (2) 우선순위에서 3순위로 선정된 '인권보장'은 가중치 분석에서는 5순위로 평가됨
- (3) 가중치 분석에서 3순위로 선정된 '투명성'은 10점 척도에서는 7순위로 낮게 평가됨
- (4) 도표의 적색 음영표시는 적용 항목의 경계선을 나타냄
- (5) 민간 측면에서 적용 중인 인공지능 윤리원칙 중 국방 인공지능 윤리원칙에 적용이 가능한 항목은 '책임성', '안전성', '침해금지', '투명성', '인권보장', '데이터관리'로 분석됨.

다) 미 국방성 인공지능 윤리원칙 항목에 대한 가중치 산출 결과

항목	가중치	내용
책임성	0.235	적절한 수준의 주의와 판단력을 지속적으로 행사 (개발, 배치, 사용에 대한 책임도 유지)
공평	0.056	AI 성능에 있어 비의도적인 편향이 최소화 되도록 신중하게 조치
추적 가능성	0.112	관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발 전 과정, 운용적방법을 적절하게 이해할 수 있도록 개발
신뢰성	0.211	인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
통제 가능성	0.385	의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있 는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

(1) 가중치 분석결과 『통제가능성 > 책임성 > 신뢰성 > 추적가능성 > 공평성』 순으로 중요도가 분석되었으며, 『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단됨.

6.4.3. 국방 인공지능 윤리 기준(안) 제언

6.4.3.1. 미 국방성 인공지능 윤리원칙 가중치 분석(AHP) 결과

가) 『통제가능성 > 책임성 > 신뢰성 > 추적가능성 > 공평성』 순으로 중요도가 분석되었으며, 『공평성』은 신뢰성에 포함될 수 있을 것으로 판단됨.

나) 즉, 『공평성』은 신뢰성에 포함될 수 있는 항목으로 판단되어 추후 전문가 의 견수렴 및 『신뢰성』의 정의 설정시 현 미 국방성 인공지능 윤리원칙 『공평성』에 정의 되어 있는 편향성 부분을 반영하는 방안 등 토론(포럼, 세미나 등)을 통해 선정

6.4.3.2. 민간 측면에서 도출된 항목 중 미 국방성 인공지능 윤리원칙 적용 가능 항목

가) 우선순위 및 가중치 분석결과 1·2·4 순위로 동일하게 분석된 핵심 키워드 『안전성, 책임성, 침해금지』는 군사용 인공지능 윤리 기준에 반영 건의

나) 민간 측면에서 우선순위 5순위, 가중치 분석에서 6순위로 평가된 『데이터 관리』는 군사용 인공지능 윤리 기준에 반영 건의

다) 민간 측면에서 우선순위 3순위, 가중치 분석에서 5순위로 평가된 『인권보장』 핵심 키워드는 군사적 활용 측면에서 적용 여부는 추가 검토가 필요(포럼, 세미나 등)

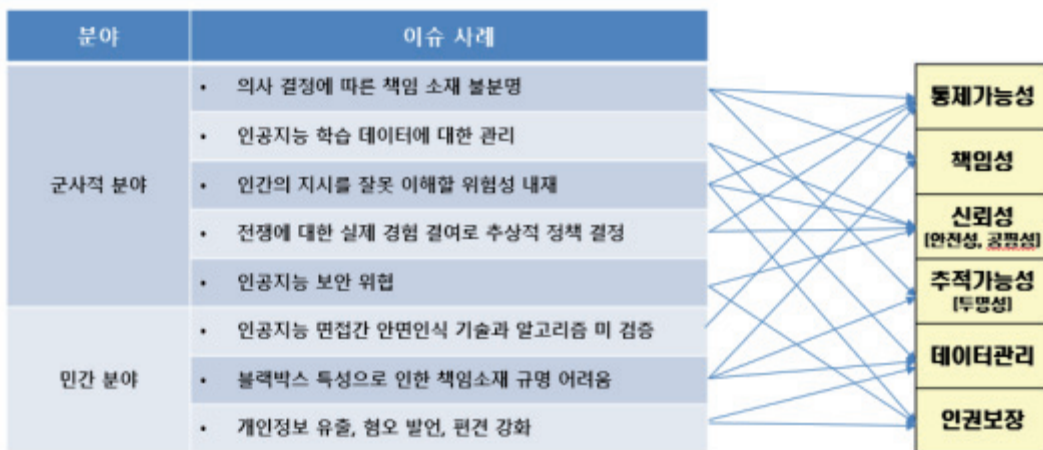
- 라) 『안전성』 → 미 국방성 인공지능 윤리원칙의 『신뢰성』에 포괄적으로 적용 가능
- 마) 『침해금지, 인권보장』 → 미 국방성 인공지능 윤리원칙의 『통제 가능성』에 포괄적으로 적용 가능
- 사) 『투명성』 → 미 국방성 인공지능 윤리원칙의 『추적 가능성』에 포괄적으로 적용

6.4.3.3. 국방 인공지능 윤리 기준 제언

- 통제가능성
 - 군사용 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생 시 사용자가 그 작동을 제어 또는 정지할 수 있는 기능을 갖추어야 한다.
 - 이상 징후 발생시 사용자가 쉽게 인지할 수 있어야 하며 이상 징후에 대한 경고 수준 설정을 할 수 있는 기능을 갖추어야 한다.
 - 인간의 의사 결정에 따라서 군사용 인공지능 시스템을 통제할 수 있어야 한다.
 - 인간의 통제를 벗어날 수 있는 잠재적 위험 요소 식별을 위해 노력하여야 한다.
- 책임성
 - 군사용 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임 소재를 명확히 할 수 있어야 한다.
 - 군사용 인공지능의 사용, 개발, 배포, 결과에 대한 책임을 정의할 수 있어야 한다.
 - 인공지능 학습을 위해 수집·저장된 데이터에 대한 관리 책임을 정의할 수 있어야 한다.
- 신뢰성(안전성, 공평성)
 - 외부 공격에 대한 회복성 등 시스템의 안정성을 보장해야 한다. 즉, 군사용 인공지능 시스템의 총수명주기간 성능의 보장, 안전, 보안 등 규정된 범위 내에서 시험되고 보증 되어야 한다.
 - 군사용 인공지능의 알고리즘에 대한 공정성과 투명성 보장을 통해 공정한 의사결정(편향성 최소화)을 유도할 수 있어야 한다.
- 추적 가능성(투명성)
 - 군사용 인공지능 알고리즘에 대한 의사결정 과정의 투명성 보장해야 한다. 즉, 군사용 인공지능 시스템은 의사결정 과정을 설명할 수 있도록 보장해야 한다.
 - 군사용 인공지능 시스템이 실패, 오작동 일으킬 경우에는 그 이유를 확인할 수 있어야 한다.
 - 군사용 인공지능은 설계로부터 전력화되는 모든 개발 과정을 이해관계자와 사용자에게 충분한 정보를 제공할 수 있어야 한다.

- 데이터 관리
 - 군사용 인공지능 개발 및 활용 간 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
 - 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발해야 한다.
 - 군사용 인공지능의 편향성(과대·과소 적합) 방지를 위한 데이터 품질 및 위험 관리 프로세스를 마련해야 한다.
- 인권보장
 - 국제인도법을 준수하며 국제법에 명시된 권리를 보장해야 한다.
 - 인간을 수단으로 취급되지 않아야 하며, 기본적 권리와 자유를 침해하지 말아야 한다.

✓ 군사용 인공지능 무기체계의 부작용 사례와 국방부 인공지능 윤리 기준(안)의 관계



붙임 1

국방 인공지능 윤리 기준 정립 연구 설문

[설문 배경]

국방부는 「국방 인공지능의 윤리 기준」을 정립하고자 정책연구를 추진하고 있습니다.

이에, 「대한민국 인공지능 윤리 기준(2020)」과 「미 국방성 AI 윤리원칙(2020)」을 참고하여 「국방 인공지능 윤리 기준」에 포함되어야 할 내용에 대해 국방 인공지능 관계자들의 의견을 수렴하여 연구에 반영하고자 합니다.

본 설문을 통하여 작성된 내용은 ‘국방 인공지능 윤리 기준 정립 연구’의 참고자료로만 활용할 것임을 알려 드립니다.

※ 설문 관련 문의처 : 국방부 데이터정책과 행정사무관
☎ 군) 900-5935

우선, 설문 작성자님의 정보를 입력하여 주시기 바랍니다.

▶ 귀하의 소속기관은 어디인가요?*

▶ 귀하의 소속 부서는 어디인가요?*

▶ 귀하의 담당 직책은 무엇인가요? (○○과장, ○○담당 등)*

- ▶ 인공지능 윤리 기준 정립과 직접적인 연관이 있는 부서에서 근무하고 계신가요?*
- 그렇다
- 아니다

- ▶ 귀하는 인공지능과 연관된 부서에서 근무하고 계신가요?*
- 그렇다
- 아니다

- ▶ 귀하는 인공지능 관련 부서에서 근무하신 적이 있으신가요?*
- 그렇다
- 아니다

- ▶ 인공지능 분야에 근무/연구/종사하신 기간은 얼마나 되시나요?*
- 0~5년
- 6~10년
- 11~15년
- 16~20년
- 20년 이상

- ▶ 국방분야의 인공지능 윤리 기준 정립의 필요성에 대해 어느 정도 공감하시나요?*
- 과거부터 인공지능 윤리 기준 정립의 필요성에 대해 인식하고 있었다.
- 최근에 인공지능 윤리 기준의 필요성에 대해 인식하게 되었다.
- 직접적인 연관은 없었으나 이번 연구 과제를 통해 인식하게 되었다.
- 아직 잘 모르겠다.
- 국방 분야 인공지능 윤리 기준은 아직 정립할 때가 아니라고 생각한다.

아래의 「대한민국 인공지능 윤리 기준」과 「미 국방성 윤리원칙」을 참고하여 주시기 바랍니다.

◎ 「대한민국 인공지능 윤리 기준(2020)」

① 인권보장

- 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 권리를 보장하여야 한다.
- 인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

② 프라이버시 보호

- 인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.
- 인공지능 전 생애주기에 걸쳐 개인 정보의 오용을 최소화하도록 노력해야 한다.

③ 다양성 존중

- 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영해야 하며, 성별·연령·장애·지역·인종·종교·국가 등 개인 특성에 따른 편향과 차별을 최소화하고, 상용화된 인공지능은 모든 사람에게 공정하게 적용되어야 한다.
- 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

④ 침해금지

- 인공지능을 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안 된다.
- 인공지능이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

⑤ 공공성

- 인공지능은 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- 인공지능은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다.
- 인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.

⑥ 연대성

- 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다.
- 인공지능 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다.
- 윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

⑦ 데이터 관리

- 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 한다.
- 데이터 수집과 활용의 전 과정에서 데이터 편향성이 최소화되도록 데이터 품질과 위험을 관리해야 한다.

⑧ 책임성

- 인공지능 개발 및 활용과정에서 책임주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- 인공지능 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다.

⑨ 안전성

- 인공지능 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능 활용 과정에서 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 갖추도록 노력해야 한다.

⑩ 투명성

- 사회적 신뢰 형성을 위해 타 원칙과의 상충관계를 고려하여 인공지능 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능 기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

◎ 「미 국방성 윤리원칙(2020)」

① 책임성(responsible)

- 적절한 수준의 주의와 판단력을 지속적으로 행사(개발, 배치, 사용에 대한 책임도 유지)

② 공평 (equitable)

- AI 성능에 있어 비의도적인 편향이 최소화되도록 신중하게 조치

③ 추적가능 (traceable)

- 관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발

④ 신뢰 (reliable)

- AI는 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증

⑤ 통제가능성 (governable)

- 의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬 수 있도록 설계

「대한민국 인공지능 윤리 기준」과 「미 국방성 윤리원칙」에서 중복된 내용을 제외하면 아래와 같습니다. 이 중 10개를 선택하여 우선순위별로 기록하여 주시기 바랍니다. 단, 고려해 주실 사항은 '국방'에 적용할 인공지능 윤리 기준이므로 민간단체나 사회적으로 중요시하는 관점과 '국방'의 시각은 다소 다를 수 있음을 유의하셔서 작성하여 주시기 바랍니다.

- ① 인권보장
- ② 프라이버시 보호
- ③ 다양성 존중
- ④ 침해금지
- ⑤ 공공성
- ⑥ 연대성
- ⑦ 데이터 관리
- ⑧ 책임성
- ⑨ 안전성
- ⑩ 투명성
- ⑪ 공평 (equitable)
- ⑫ 추적가능 (traceable)
- ⑬ 신뢰 (reliable)
- ⑭ 통제가능성 (governable)

- ▶ 이외에 추가해야 할 윤리 기준이 있다면 제시하여 주시기 바랍니다. (선택사항)
기준에 대한 설명도 곁들여 주시기 바랍니다.
또한 아래 10개의 우선순위에 포함하여 작성하셔도 됩니다.

▶ 첫번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 두번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 3번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?

▶ 4번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 5번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 6번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 7번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 8번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 9번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 10 번째 우선순위는 무엇입니까? 그리고 그 이유는 무엇입니까?*

▶ 국방분야의 인공지능 윤리 기준은 몇 개가 적당하다고 생각하십니까?*

- ☐ 3개 이하
- ☐ 4~5개
- ☐ 6~8개
- ☐ 9~10개
- ☐ 11개 이상

✓ 위에서 선택하신 항목이 중복되지는 않는지 확인해 주시기 바랍니다. 설문 응답에 감사드립니다.

붙임 2



안녕하십니까?

본 설문은 『국방부 데이터정책과』와 『건양대학교』에서 연구수행하는 “국방 데이터·인공지능 윤리 기준 정립 연구”와 관련하여 관련분야 전문가분들의 소중한 경험과 식견으로 의견을 수렴하고자 합니다.

의견수렴 사항은 국방 인공지능 윤리 기준 정립을 위한 원칙(핵심 키워드) 설정과 키워드에 대한 가중치 설정을 위해 의견을 수렴하고자 하는 설문 내용입니다.

설문자의 소중한 의견은 국방 인공지능 윤리 기준 정립을 위한 연구수행 방안 및 설정에 큰 도움이 될 것입니다.

설문에 응답 부탁드립니다.

감사합니다.

* 본 설문은 국방 인공지능 윤리 기준 원칙 핵심 키워드별 가중치 설정을 목적으로 하며, 응답한 귀하의 의견에 대해서는 통계법 제13조에 의거하여 그 비밀이 철저히 보장되오니, 바쁘시더라도 성실한 답변을 해 주시어 귀하의 고견을 꼭 반영할 수 있도록 해 주시면 감사하겠습니다.

2023년 9월

1) AHP (Analytic Hierarchy Process) 분석 목적

□ 정의

- AHP 분석은, 우선순위 도출을 위해 두 Factor 간 쌍대비교 행렬화 연산을 이용한 가중치 산정 기반 의사결정 도구로, 평가항목들을 계층적으로 구성하고 쌍대비교를 통해 문제를 해결하는 방법임

※ 사티(Saaty) 교수가 1970년 대에 제안, 고유벡터 이론을 기반으로 하며, 일관성 있는 비교 평가로 대안을 선정하게 된다.

□ 목적

- 인공지능 기술의 급속한 발전으로 제조·의료·교통·환경·교육 등 산업 전 분야에 인공지능이 활용·확산되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 인공지능 윤리 이슈가 새롭게 대두
- 이에 따른 국방 인공지능 실현을 위한 윤리적 측면의 원칙(핵심 키워드)를 설정하고자 함.

□ 설문 대상

- 인공지능 관련부서 근무(업무) 유경험자
 - 정책부서, 각 군 관련부서, 교육 및 연구기관 등 전문가 집단 선정
- 분석 절차
 - 국방 인공지능 윤리 기준 정립에 따른 원칙(핵심 키워드) 도출
 - ※ 대한민국 인공지능 윤리 기준, 미 국방성 윤리원칙, 아실로마 인공지능 원칙, UNESCO AI 윤리, 국내 민간 부문 윤리 헌장 등 관련 문헌을 통해 주요 사항에 대해서 도출
 - 인공지능 윤리 기준 정립을 위한 원칙 즉, 핵심 키워드별 가중치 분석을 위해 AHP 기법을 적용하여 키워드별 상대적 가중치를 결정
 - 설문조사는 과제 제기기관에 의뢰 / 실시하여 결과를 도출
 - 도출한 가중치 결과를 통해서 국방 인공지능 윤리 기준 정립을 위한 원칙과 핵심 키워드 설정의 기초자료로 활용

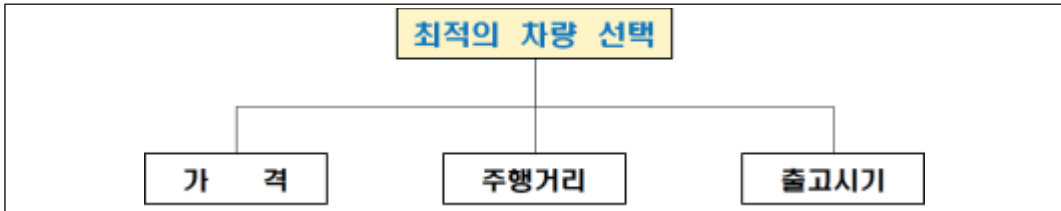
AHP 설문 요령(예시)

□ 중고자동차 구매

A씨는 중고자동차를 구매하려고 합니다. A씨의 중고자동차 구매 기준은 가격과 주행거리, 그리고 출고시기이며, AHP 기법을 적용하고자 합니다.

□ AHP 설문

○ 효과요소 계층구조



○ 최적의 차량 선택을 목표로 하고, 이를 결정하는 주요 요소로 가격, 주행거리, 출고시 기간의 상대적인 중요도를 구하고자 합니다.

요소		절대 중요		매우 중요		중요		약간 중요		같 다		약간 중요		중요		매우 중요		절대 중요		요소	
Ⓐ	가 격	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	Ⓑ	주행거리	
Ⓐ	가 격	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	Ⓒ	출고시기	
Ⓑ	주행거리	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	Ⓒ	출고시기	

① ①과 ② 비교 : ①가 “매우 중요” / ① ≫ ②

② ①과 ③ 비교 : ①가 “약간 중요” / ① > ③

③ ②와 ③ 비교 : ③가 “중요” / ② ≪ ③

※ ■에 표시되어야 응답의 일관성이 유지가 됨

국방 인공지능 윤리 기준 설정에 따른 원칙(핵심 키워드)

원칙(핵심 키워드)		정의
민간 측면	인권보장	- 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중 - 민주적 가치와 국제 인권법 등에 명시된 권리를 보장
	프라이버시 보호	- 인공지능을 개발하고 활용하는 전 과정에서 개인정보 보호 - 인공지능 전 생애주기에 걸쳐 개인정보의 오용을 최소화
	침해금지	- 인공지능은 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안됨 - 인공지능의 위험과 부정적 결과에 대한 대응 방안 마련
	다양성 존중	- 인공지능 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영 - 성별·연령·장애·지역·인종·종교·국가 등 개인 특성에 대한 편향과 차별 최소화
	공공성	- 인공지능은 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익 추구 - 인공지능은 긍정적, 순기능을 극대화(역기능을 최소화)
	연대성	다양한 집단 간의 관계 연대성을 유지, 다양한 주체들의 공정한 참여 기회를 보장
	데이터 관리	- 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용(목적외 활용 금지) - 데이터 수집과 활용의 전 과정에서 데이터 편향성 최소화, 데이터 품질과 위험 관리
	책임성	- 인공지능 개발 및 활용과정에서 책임 주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력 - 인공지능 설계, 개발자, 서비스 제공자, 사용자 간의 책임소재 명확
	안전성	- 인공지능 개발 및 활용 전 과정에서 잠재적 위험 방지(안전 보장) - 명백한 오류 또는 침해가 발생할 때 사용자가 그 작동을 제어
국방 측면	투명성	- 적합한 수준의 투명성과 설명 가능성을 높여야 함. - 인공지능 활용 과정에 발생할 수 있는 위험 등 사전에 고지 가능
	책임성	적절한 수준의 주의와 판단력을 지속적으로 행사(개발, 배치, 사용에 대한 책임 유지)
	공평	인공지능 성능에 있어 비의도적인 편향이 최소화 되도록 조치
	추적 가능성	관련자들이 투명하고 감시 가능한 방법, 데이터의 근거 등을 포함한 기술, 발전 과정, 운용적 방법을 적절하게 이해할 수 있도록 개발
	신뢰성	인공지능은 분명하고 명확하게 사용되어야 하며, 이러한 성능의 효과, 안전, 보안은 전체 수명주기 동안 규정된 사용 범위 내에서 시험되고 보증
	통제 가능성	의도된 기능을 다하도록 설계하되, 비의도적인 결과를 탐지하고 회피할 수 있는 성능과 비의도적인 작동을 하는 경우 중단시킬수 있도록 설계

2) 국방 인공지능 윤리 기준 정립에 따른 원칙(핵심 키워드) 가중치 평가

□ 상대적 중요도

- 국방 인공지능 윤리 기준 설정을 위한 국방 측면 원칙(핵심 키워드) 5개 항목 중 상대적으로 중요하다고 생각되는 요소에 √ 표시를 해주시기 바랍니다.

요소	절대 중요		매우 중요		중 요		약간 중요		같 다		약간 중요		중 요		매우 중요		절대 중요	요소
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	공평
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	추적가능
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	신뢰
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	통제가능
공평	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	추적가능
공평	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	신뢰
공평	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	통제가능
추적가능	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	신뢰
추적가능	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	통제가능
신뢰	⑨	⑧	⑦	⑥	⑤	④	③	②	①	②	③	④	⑤	⑥	⑦	⑧	⑨	통제가능

- 민간 측면의 원칙(핵심 키워드) 중 국방 측면에 반영되어야 하는 원칙을 우선순위 순으로 표시 (1부터 10까지 숫자로 작성해 주세요)

원칙(핵심 키워드)	정의	우선순위
인권보장	• 인공지능의 개발과 활용은 모든 인간에게 동등하게 부여된 권리 존중	
프라이버시 보호	• 인공지능을 개발하고 활용하는 전 과정에서 개인정보 보호	
침해금지	• 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안됨	
다양성 존중	• 개발 및 활용 전 단계에서 사용자의 다양성과 대표성을 반영	
공공성	• 개인적 행복 추구 뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익 추구	
연대성	• 다양한 집단 간의 관계 연대성을 유지, 다양한 주체들의 공정한 참여 기회를 보장	
데이터 관리	• 개인정보 등 각각의 데이터를 그 목적에 부합하도록 활용(목적외 활용 금지)	
책임성	• 인공지능 개발 및 활용과정에서 책임 주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력	
안전성	• 인공지능 개발 및 활용 전 과정에서 잠재적 위험 방지	
투명성	• 적합한 수준의 투명성과 설명 가능성을 높여야 함.	

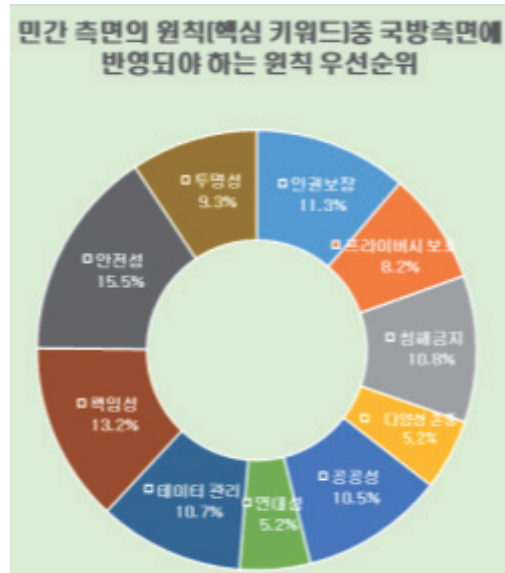
- 국방 인공지능 윤리 기준 설정을 위한 민간 측면 원칙(핵심 키워드) 10개 항목 중 상대적으로 중요하다고 생각되는 요소에 √ 표시를 해주시기 바랍니다.

요소	절대 중요	매우 중요	중요	약간 중요	같 다	약간 중요	중요	매우 중요	절대 중요	요소
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	프라이버시 보호
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	침해금지
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	다양성 존중
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	공공성
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	연대성
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
인권보장	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	침해금지
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	다양성 존중
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	공공성
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	연대성
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
프라이버시 보호	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성

요소	절대 중요	매우 중요	중 요	약간 중요	같 다	약간 중요	중 요	매우 중요	절대 중요	요소
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	다양성 존중
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	공공성
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	연대성
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
침해금지	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	공공성
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	연대성
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
다양성 존중	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
공공성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	연대성
공공성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
공공성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
공공성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
공공성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
연대성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	데이터 관리
연대성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
연대성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
연대성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
데이터 관리	⑨	⑧	⑦	⑥	⑤	④	③	②	①	책임성
데이터 관리	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
데이터 관리	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	안전성
책임성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성
안전성	⑨	⑧	⑦	⑥	⑤	④	③	②	①	투명성

□ 민간 측면의 원칙(핵심 키워드) 중 국방 측면에 반영되어야 하는 원칙을 우선순위

원칙(핵심 키워드)		우선순위(점)	순위	비율
민간 측면	인권보장	93	3	11.3%
	프라이버시 보호	68	8	8.2%
	침해금지	89	4	10.8%
	다양성 존중	43	9	5.2%
	공공성	87	6	10.5%
	연대성	43	9	5.2%
	데이터 관리	88	5	10.7%
	책임성	109	2	13.2%
	안전성	128	1	15.5%
	투명성	77	7	9.3%
	총계	825		100%



□ 상대적으로 중요 가중치 분석에 따른 신뢰도

Eigen Vector 계산			일관성 검토			
1	행합계	정규화	1	대비교행렬X Eigen Vec	①/Eigen Vector	CI
책임성	89.575	0.23552	책임성	1.200	5.094	0.098
공명	18.040	0.04743	공명	0.266	5.603	RI
추적가능성	13.004	0.03419	추적가능성	0.194	5.665	1.12
신뢰성	25.696	0.06756	신뢰성	0.366	5.415	
통제가능성	234.019	0.61530	통제가능성	3.188	5.182	CR
합계	380.335	1.00000	Consistency	Yes	0.391826	0.087
2	행합계	정규화	2	대비교행렬X Eigen Vec	①/Eigen Vector	CI
책임성	108.733	0.35327	책임성	1.859	5.262	0.097
공명	9.823	0.03191	공명	0.184	5.771	RI
추적가능성	27.540	0.08948	추적가능성	0.495	5.530	1.12
신뢰성	114.067	0.37060	신뢰성	1.989	5.368	
통제가능성	47.629	0.15474	통제가능성	0.775	5.006	CR
합계	307.791	1.00000	Consistency	Yes	0.387358	0.086
3	행합계	정규화	3	대비교행렬X Eigen Vec	①/Eigen Vector	CI
책임성	73.000	0.44042	책임성	2.205	5.007	0.005
공명	14.250	0.08597	공명	0.433	5.035	RI
추적가능성	26.167	0.15787	추적가능성	0.792	5.019	1.12
신뢰성	26.167	0.15787	신뢰성	0.792	5.019	CR
통제가능성	26.167	0.15787	통제가능성	0.792	5.019	
합계	165.750	1.00000	Consistency	Yes	0.019854	0.004
4	행합계	정규화	4	대비교행렬X Eigen Vec	①/Eigen Vector	CI
책임성	30.497	0.08911	책임성	0.482	5.411	0.107
공명	14.897	0.04353	공명	0.249	5.712	RI
추적가능성	14.897	0.04353	추적가능성	0.249	5.712	1.12
신뢰성	85.076	0.24859	신뢰성	1.244	5.006	
통제가능성	196.867	0.57524	통제가능성	3.051	5.305	CR
합계	342.234	1.00000	Consistency	Yes	0.429308	0.096

✓ CR 값 0.1보다 큰 설문조사 결과는 배제하였으며, 가중치 평균 산출은 기하평균임

참고문헌

가. 국내 단행본 및 논문

- [1] 과학기술정보통신부, 정보통신정책연구원, 2023 인공지능 윤리 기준 실천을 위한 자율 점검표(안)
- [2] 권지혜(2016), 형법 정책의 토대로서 자유주의적 후견주의에 대한 비판적 고찰: 드워킨(Gerald Dworkin)과 파인버그(Joel Feinberg)의 논의를 중심으로, 고려대 법학박사 학위논문.
- [3] 김진우(2017), 법철학 연구, 제20권 제2호, pp. 7~44.
- [4] 김나경 (역)(2002), 이상돈(엮음), 대화이론과 법, 법문사, pp. 152~154.
- [5] 김도균(2008), 권리의 문법, 박영사.
- [6] 김영평, 최병선, 신도철 (편저)(2006), 규제의 역설, 삼성경제연구소.
- [7] 김자희, 장신, 주성구(2017), 자율 로봇의 잠재적 무기화에 대한 소고 - 개념정립을 통한 규제를 중심으로, 국회입법조사처, 입법과 정책, 제9권 제3호 2017. 12.
- [8] 노태복(윝김)(2016), 왜 로봇의 도덕인가, 메디치미디어, p. 13.
- [9] 노훈, 이재욱(2001), 사이버전의 출현과 영향, 그리고 대응방향, 국방정책연구, 가을호.
- [10] 문정욱 외(2022. 12), AI윤리체계 마련, 인공지능 윤리체계 확립을 위한 정책연구, 과학 기술정보통신부.
- [11] 박혜성, 김법연, 권현영 (2021), 인공지능 통제를 위한 규제의 동향과 시사점, 정보법학, 25권, pp. 1~39.
- [12] 백종현(윝김)(2005), 윤리형이상학 정초, 아카넷.
- [13] 변순용 외, 자율주행자동차에 대한 한국형 윤리 가이드 라인 연구, 윤리연구, 제112호, 제123호.
- [14] 서병훈(윝김)(2018), 자유론, 책세상.
- [15] 심재우, 윤재왕(윝김)(2019), 법치국가와 인간의 존엄, 세창출판사.
- [16] 안예찬, 신영현, 이근호(2015), 빅데이터 플랫폼과 모니터링 시스템의 융합을 이용한 BLE 기반의
- [17] 오병선(2006), 밀의 자유론과 해악의 원리, 『수행인문학』, 제36집, pp. 31~64.
- [18] 오윤경 외(2020), 혁신과 위험관리: 사람중심 기술혁신을 위한 추진과제, ISSUE PAPER, 한국 행정연구원, p. 9.
- [19] 윤정현(2020), 국방 분야의 인공지능 활용성 제고 방안과 시사점, FUTURE HORIZON, pp. 44~52.
- [20] 양천수(2009), 법학연구, 제50권 제1호, 부산대학교, pp. 161~191.
- [21] 양천수(2014), 법철학연구, 제17권 제2호, pp. 205~242.

- [22] 양천수(2017), 법철학 연구, 제20권 제2호, pp. 45~76.
- [23] 양해림 외(2018), 공학윤리, 충남대학교 출판부 양해림, p. 55.
- [24] 이주석(2017), 도덕직관에 관한 윤리적 해명 가능성: 트롤리 딜레마의 경우, 철학논총, 제90집, pp. 363~380.
- [25] 이상돈, 정채연(2017), IT와 법 연구, 제15집, pp. 281~325.
- [26] 이진우(역)(1997), 담론윤리의 해명, 문예출판사.
- [27] 이준일(윝김)(2007), 기본권이론, 한길사.
- [28] 이중기, 오병두(2016), 홍익법학, 제17권 제2호, pp. 1~25.
- [29] 이현준, 홍영기(역)(2018), 형사법 연구, 제30권 제1호, pp. 207~234.
- [30] 정승원, 황기연(2018), 자율주행 자동차의 윤리적 선택에 따른 교통사고비용 분석, 한국 ITS학회 논문지, 제17권 제6호, pp. 224~239.
- [31] 정진규(2016), 트롤리 문제와 다원론적 규범 윤리 이론, 동서철학연구, 제81호, pp. 423~446
- [32] 정혜수(2021), 한국형 무기체계 CBM 적용 및 확대를 위한 상태기반정비 발전방향 연구, 한국산학기술학회 논문지, 22(8), pp. 631~638.
- [33] 조성은 외(2018), 인공지능시대 법제 대응과 사회적 수용성, 정보통신정책연구원.
- [34] 조현식(2018), 인공지능 자율무기체계와 미래 전쟁의 변환, 21세기정치학회보, 제28집 제1호.
- [35] 최명진 외(2022), 군사용 인공지능의 안전성 확보를 위한 단계별 윤리적 평가 및 검증요소 (기준) 마련 연구.
- [36] 최은창, 컴퓨터 알고리즘은 윤리적 전투방법을 배울 수 있을까?, MIT Technology Review,
- [37] 한국국방연구원, 군사적 인공지능의 법적(윤리적) 문제에 대한 검토, 2020.4.27. pp. 19~20
- [38] 한국인공지능법학회(2019), 인공지능과 법, 박영사, p. 57.
- [39] 한희원, 정의로의 산책(제2판), 삼영사, 2011, pp. 214~216.
- [40] 한희원(2018), 인공지능(AI) 치명적자율무기(LAWs)의 법적·윤리적 쟁점에 대한 기초 연구, 중앙법학, 20(1), pp. 325~365.

나. 국외 단행본 및 논문

- [1] ALTAI : The Assessment List for Trustworthy Artificial Intelligence
- [2] Campaign to Stop Killer Robots. "High-Level Concerns on Killer Robots at UN."
- [3] Campaign to Stop Killer Robots, Country Views on Killer Robots, 14 April, 2016.

- [4] Czajkowski, M., & Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. *Expert Systems with Applications*, 137, pp.392~404.
- [5] *Deadly Decisions - 8 objections to killer robots*, 2016, p.10.
- [6] D. N. Hammond, *Autonomous Weapons and the Problem of State Accountability*, *Chicago Journal of International Law*, Vol. 15(2), 2015.
- [7] Fazl Barez, "Interactive Intelligence: Human-In-The-Loop Intelligence," *TheDataLab*, April 22, 2019.
- [8] Firlej, M., & Taeihagh, A. (2021). Regulating human control over autonomous systems. *Regulation & Governance*, 15(4), pp.1071~1091.
- [9] *Future of life institute, Asilomar AI Principles*, 2017.
- [10] Gunter. Teubner(1984), *Verrechtlichung - Begriffe, Merkmale, Grenzen, Auswege*, in: F. Kübler (Hrsg.), *Verrechtlichung von Wirtschaft, Arbeit und sozialer Solidarität (Frankfurt/M.)*, S. 312 ff.
- [11] *Getting to grips with military robotics*, *The Economist*, 25 January 2018.
- [12] *Human Rights Watch and the Harvard Law School International Human Rights Clinic, Mind the Gap: The Lack of Accountability for Killer Robots*, April 2015.
- [13] I. Kant(1797), *Metaphysik der Sitten, Einleitung in die Rechtslehre*, §B, S. 337.
- [14] Jacob Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots", *American Civil Liberties Union(ACLU)*, 2018. 7. 26.
- [15] J. Sydney, Jr. Freedberg, *Secretary Of Drones: Mabus Creates DASN for Unmanned*, *Breaking Defence*, 2015.
- [16] Kim, J. H., Lim, S., Kim, I. K., Cho, H. S., & No, B. K. (2013). Technical trends of cyber security with big data. *Electronics and Telecommunications Trends*, 28(3), pp.19~29.
- [17] Muslim, M. A., & Dasril, Y. (2021). Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. *International Journal of Electrical & Computer Engineering* (2088-8708), 11(6).
- [18] M. Schmitt, 'Autonomous Weapon Systems and International Humanitarian Law: A reply to critics', *Harv Nat'l Sec J* at 2, 2012.
- [19] Musandu, N. N. (2016). *Doctor of Philosophy Thesis in Military Informatics (OpenPhD): Lethal Autonomy of Weapons is Designed and/or Recessive*.

- [20] Ministry of Defence of the UK, Joint Doctrine Note 2/11: The UK Approach to Unmanned Aircraft Systems, 30. 3. 2011.
- [21] Nyagudi, Nyagudi Musandu(2016-12-09). "Doctor of Philosophy Thesis in Military Informatics: Lethal Autonomy of Weapons is Designed and/or Recessive"
- [22] Park, H. K. (2013). Types and Information Security Technology on Cyber Warfare. Review of Korea Contents Association, 11(4), pp.41~44.
- [23] Pentagon examining the 'killer robot' threat". Boston Globe. 30 March 2016.
- [24] Protocol I additional to the Geneva Conventions, 1977, arts. 86 (2) and 87.
- [25] Robinson, M, Jones, K., & Janicke, H. (2015). Cyber warfare: Issues and challenges. Computers & security, 49, pp.70~94.
- [26] Rahman, A.(2020), Software Defect Prediction Using Rich Contextualized Language Use Vectors (Doctoral dissertation, Queen's University (Canada).
- [27] Russell, S. (2015). Take a stand on AI weapons. Nature, 521(7553), pp.415~416.
- [28] Russell, S., Hauert, S., Altman, R., & Veloso, M. (2015). Ethics of artificial intelligence. Nature, 521(7553), pp.415~416.
- [29] Russell, Stuart(27 May 2015). "Take a stand on AI weapons". International weekly journal of science. 521.
- [30] Rome Statute of the International Criminal Court, Article 8(War crimes) 2(b).
- [31] Simpson, Thomas W; Muller, Vincent C. (2016). "Just war and robots' killings". Philosophical Quarterly.
- [32] Singh, J., Walden, I., Crowcroft, J., & Bacon, J. (2016). Responsibility & machine learning: Part of a process. Available at SSRN 2860048.
- [33] United Nations Geneva. "Position Paper, Submitted by China," CCW/GGE.1/2018/WP.7, para.3.
- [34] United Nations Geneva. "Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapon Systems in the Context of the Purposes and Objectives of the Convention, Submitted by the Russian Federation," CCW/GGE.1/2018/WP.6, para.2(a).
- [35] US says Russia 'developing' undersea nuclear-armed torpedo". CNN. 2018.
- [36] U.S. Department of Defense Directive, No. 3000.09, 21 November 2012, pp. 13~14.
- [37] US Department of Defense(2012). "Directive 3000.09, Autonomy in weapon systems" p. 2; Gubrud, Mark (April 2015). "Semi-autonomous and on their own: Killer robots in Plato's Cave". Bulletin of the Atomic Scientists.

- [38] U.S. Delegation Opening Statement as delivered by M. Meier, CCW Informal Meeting of Experts on Lethal Autonomous Weapons Systems, 13 April 2015.
- [39] Women's International League for Peace and Freedom(WILPF), Statement to the CCW Meeting of Experts on Lethal Autonomous Weapon Systems, 2014.
- [40] Yannakakis, G. N., & Togelius, J. (2018). Artificial intelligence and games(Vol. 2, pp. 2475-1502). New York: Springer.

다. 인터넷

- [1] <https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1:primaryr6>
- [2] <https://ai.google/principles>
- [3] <https://www.stopkillerrobots.org/>
- [4] <https://www.joongang.co.kr/article/23875682#home>
- [5] https://biz.chosun.com/site/data/html_dir/2018/04/09/2018040900258.html
- [6] <https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>
- [7] <http://www.theguardian.com/technology/2013/feb/23/stop-killer-robots>
- [8] Lewis, M. W. "The Misleading Human Rights Watch and Amnesty International Reports on US Drones."
<http://opiniojuris.org/2013/11/08/misleading-human-rights-watch-amnesty-international-reports-u-s-drones>
- [9] <http://www.stopkillerrobots.org/the-solution/>
- [10] <http://www.hrw.org/topic/arms/killer-robots>.

유 의 사 항

1. 본 연구보고서 내용은 연구진의 개인적인 견해로서 국방대학교 국가안전보장문제연구소의 공식입장과 다를 수 있습니다.
2. 본 연구보고서는 정책입안시 참고자료로만 활용하고 타 기관에 불필요한 자료유출을 삼가 주시기 바랍니다.

ISSN 1976-5967

Volume 16

2023. 12



Military Science & Technology Annual Report



Research Institute for National Security Affairs
Korea National Defense University