EU 인공지능(AI) 윤리 가이드라인 연구

김창화(Kim, Chang-Hwa)[1)]

Ⅰ. 서언

- 인공지능(AI)은 인간에게 많은 혜택을 가져다주기도 하지만, 잠재적 위험 발생의 가능성도 매우 큰 것으로 여겨짐
- 2017년 글로벌 리스크 보고서(Global Risk Report)에 따르면, AI와 로봇공학이 12개의 새로운 기술 중 편익과 위험이 모두 가장 큰 것으로 발표되었음
- Potential Harms Caused by AI Systems: 편결과 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
- 세계 각국은 AI에 의한 부작용을 방지하기 위해 AI 윤리와 관련된 대책을 마련하고 있음
- 유럽연합 집행위원회는 ‘신뢰할 수 있는 AI를 위한 윤리 가이드라인’을 발표하였음
- 미국은 ‘AI의 윤리적 발전을 위한 결의’을 마련하고, AI의 보안과 프라이버시 등에 대한 윤리기준을 구체화하고 있음
- 일본은 정부 주도 ‘인간 중심 AI 사회 원칙 검토회의’를 통해 7대 윤리기준을 제정함
- 중국은 국가 차세대 AI 관리 특별위원회에서 ‘차세대 AI 관리 원칙’을 발표하고 AI 시스템 개발의 프레임과 액션 가이드라인을 제시함

- 우리나라는 AI 국가전략에서 사람중심의 AI 구현을 제시하고, 안전한 AI 사용을 위한 AI 역기능 방지와 AI 윤리 정립을 제안한 바 있음

인공지능(AI) 국가전략 발표

[9] 역기능 방지 및 AI 윤리체계 마련

❶ AI 기반 **사이버침해 대응체계** 고도화(’20~)

❷ **딥페이크*** 등 신유형의 **역기능 대응**을 위한 범부처 협업체계 구축(‘20)

* AI 기반 영상 합성기술 또는 그 영상. 신시장 창출과 동시에 명예훼손 등 부작용

1) 한밭대학교 공공행정학과 부교수, 법학박사(S.J.D.)

도 우려

❸ AI 신뢰성·안전성 등을 검증하는 **품질관리체계** 구축 추진(’ 20~)

❹ OECD 등 **글로벌 규범**에 부합하는 **AI 윤리기준 확립**(’ 20) 및 **AI 윤리교육 커리큘럼*** 개발·보급(’ 21~)

* (학생·이용자) AI와 생명윤리, 개인정보보호 / (개발자) 윤리적 AI 설계, 정보보안 등

❺ **이용자 보호**를 위한 중장기적 **정책 수립 지원체계** 마련

- 국내 AI 윤리 헌장은 정부와 공공기관이 제정한 윤리 헌장 5개와 비영리기관과 카카오가 마련한 윤리 헌장 2개 등 총 7개가 있음
- 국내 윤리 헌장 중에는 지능정보사회 윤리 헌장이 있으나, 모든 기술에 적용되는 윤리기준이라고 볼 수 없고, 그 외 헌장은 각기 AI 사용 분야별로 제정돼 있어 AI 전반에 관한 윤리기준이 미흡한 상태임
- 우리나라도 AI의 윤리적 사용을 위한 표준화된 기준 마련 등 개선이 요구됨

Ⅱ. EU 인공지능 윤리 가이드라인

1. 개요

- 2018년 유럽위원회(EC)는 윤리적이고, 안전하며, 최첨단의 AI를 위한 유럽에서의 AI 비전을 제시하였음
- 유럽위원회는 비전의 실행을 돕기 위하여 AI에 대한 고위 전문가 그룹(High-Level Expert Group, HLEG)을 설립하였고, 이들에게 윤리 가이드라인을 만들도록 하였음
- 윤리 가이드라인은 AI가 사회를 의미 있게 변화시킬 수 있는 잠재력이 있지만, AI 그 자체가 목적이 되어서는 안 되고, 인간의 번영을 위한 수단이 되어야 하므로 윤리 가이드라인은 인간 중심적이어야 한다고 하였음
- 또한, AI는 적정하게 다루어져야만 하는 위험들을 일으킬 수 있으므로 AI의 혜택을 최대화하는 동시에 그 위험을 막기 위해 노력해야 한다고 하였음
- 빠른 기술적 변화의 환경에서, 신뢰는 사회, 공동체, 경제 그리고 지속 가능한 개발의 기반이 되며, 인간과 공동체는 신뢰성을 달성하기 위한 명확하고 포괄적인 체계가 마련될 때만 기술의 개발과 적용에 대한 확신을 가질 수 있으므로, 신뢰할 수 있는 AI가 기본적 목표가 됨
- 신뢰성은 사람과 사회가 AI 시스템을 개발하고, 배치하고 이용하기 위한 선결조건임
- 명백하게 신뢰할 만한 AI 시스템이 없고 그들 뒤에 인간이 없다면, 원하지 않는 결과들이 뒤따를 수 있고 사회·경제적 혜택의 실현을 막음으로써 그 활용이 저해될 수 있음
- 유럽이 그러한 혜택의 달성을 돕기 위해, 유럽의 AI 비전은 신뢰할 수 있는 AI를 보장하고 확장하는 것임
- AI 시스템의 개발, 배치, 이용에서의 신뢰는 기술의 본질적인 특성뿐만 아니라 AI의 적용과 관련된 사회-기술적 시스템의 질과 관련되며, AI 시스템의 단순한 요소가 아니고 신뢰를 불러일으키거나 일으키지 않을 수 있는 총체적 환경에서의 시스템임
- 신뢰할 수 있는 AI를 얻으려고 노력하는 것은 이런 이유로 AI 그 자체의 신뢰성과 관련될 뿐만 아니라 모든 행위자들의 신뢰성과 전주기 내내 시스템의 사회-기술적 환경의 부분인 절차들을 포함하기 때문에, 총체적이고 체계적인 접근을 요구함

- 신뢰할 수 있는 AI는 3가지 구성요소를 가지며, 그것은 시스템의 전주기를 통해 만족되어야 함
  ⅰ. 신뢰할 수 있는 AI는 모든 적용 가능한 법과 규칙을 따라야 하기에, 합법적이어야 함
  ⅱ. 윤리적 원칙과 가치의 준수를 보장하기에, 윤리적이어야 한다
  ⅲ. 비록 좋은 의도를 가졌을지라도 AI 시스템이 의도하지 않은 피해를 일으킬 수 있어, 기술적·사회적 관점에서 강건해야 함
- 이러한 3가지 구성요소들은 신뢰할 수 있는 AI를 달성하기 위한 필요조건이지만, 그것 자체로 충분요건은 되지 않으며, 3가지 구성요소들은 이상적으로는 조화롭게 작동해야 하지만, 실제로는 기존법의 범위와 내용이 윤리적 규범과 맞지 않을 수도 있는 등 이들 사이에 충돌이 생길 수 있고, 그렇다면 그것들을 조정하기 위한 노력이 필요함
- 신뢰성 접근 방식은 AI 시스템과 관련된 사람들이 그들의 디자인, 개발 그리고 이용이 합법적이고, 윤리적이고 강건하다는 것을 믿을 수 있는 근거를 제공함으로써 "책임 있는 경쟁력"을 가능케 하는 데 필수적임
- 신뢰성을 담보함으로써 유럽의 개인들은 AI 시스템의 혜택을 완전히 얻을 수 있고, 잠재적 위험으로부터 보호하는 조치들이 취해질 것을 알고 안심할 수 있음

2. 신뢰할 수 있는 AI의 체계

(1) 신뢰할 수 있는 AI의 3가지 구성요소

- AI 시스템은 법이 없는 세상에서 작동하지 않으며, 국가, 유럽 또는 국제적 수준에서 법적 구속력 있는 많은 규정이 AI 시스템의 개발, 배치, 이용에 이미 적용되거나 관련되고 있음[1)]
- 건강 관리 분야에서 의학 장치 규정과 같이 특정한 AI에 적용되는 다양한 규정들도 존재함
- 가이드라인은 첫 번째 구성요소인 합법적 AI를 다루지 않고, 두 번째와 세 번째 요소들을 촉진하고 보장하는 가이드를 제공함
- 2개의 요소는 기존의 법들로부터 영향을 받지만, 그들의 완전한 실현은 기존 법적 의무를 능가할 수도 있음
- 본 가이드라인은 기존 법적 규범과 요건에 영향을 미치지 않으며, 법적 권리와 의무를 만들지 않지만, AI 시스템을 개발하고, 배치하고 이용하는 것과 관련된 절차와 행위들에 적용되는 모든 법적 권리들과 의무들이 반드시 지켜져야 한다

1) 법적 출처들은 유럽연합의 1차법으로서 유럽연합 조약과 기본권 헌장, 2차법으로서 개인정보 보호기본법(GDPR), 제품책임지침, 비개인 데이터의 자유 흐름에 관한 규정, 반차별 지침, 소비자법 그리고 안전보건지침, UN 인권 조약과 유럽 인권 협약, 그리고 여러 유럽연합 회원국들의 법들이 있다.

는 가정하에 진행됨

- 법은 항상 기술적 개발의 속도를 따라갈 수 있지 않으며, 때로는 윤리적 규범과 조화되지 못하거나 특정 이슈들을 다루는 데 적합하지 않을 수도 있음
- AI 시스템이 신뢰할 수 있으려면 윤리적 규범과의 일치를 보장함으로써 윤리적이어야 함
- 윤리적 목적이 보장될지라도, 개인과 사회는 AI 시스템이 의도되지 않은 피해를 일으키지 않을 것에 대해 자신감이 있어야 하고, 그러한 시스템은 안전하고, 확실하며, 믿을 수 있는 방법으로 수행되어야만 하고, 의도되지 않은 악영향을 막을 수 있는 안전장치가 예측될 수 있어야 함(강건함)
- 이것은 기술적 측면(지역이나 전주기 단계에서 적용되는 것과 같이 주어진 환경에서 적정하게 시스템의 기술적 강건함을 보장하는 것)과 사회적 측면(시스템이 운영되는 사정과 환경을 충분히 고려해서) 양자에서 모두 필요함
- 윤리적이고 강건한 AI는 따라서 밀접하게 관련되어 있고 서로를 보완함

(2) 신뢰할 수 있는 AI의 근거

- 신뢰할 수 있는 AI의 근거는 기본권에 근거를 두고 있으며, 윤리적이고 강건한 AI를 보장하기 위해 준수되어야 할 4가지 윤리 원칙들로 나타내짐
- 근거는 주로 윤리학 분야에 큰 비중을 두고 있으며, AI 윤리학은 응용 윤리학의 하위 분야이며, AI의 개발, 배치, 이용으로 제기된 윤리적 이슈들에 초점이 맞추어져 있음
- 그 본질은 삶의 질이나 민주 사회에 필요한 인간의 자율성과 자유 측면에서 AI가 어떻게 개인의 삶에 대한 문제를 야기하거나 진행시키는지를 확인하는 것임
- AI 기술에 대한 윤리적 성찰은 다양한 역할을 수행할 수 있음: 첫째, 가장 기본적인 수준에서 개인과 집단을 보호해야 할 필요성에 대한 성찰을 자극할 수 있음; 둘째, 다가올 EU 아젠다 2030에 확고하게 들어갈 UN의 지속 가능한 개발 목표를 성취하는데 돕는 것들과 같이, 윤리적 가치를 촉진하기 위해 노력하는 새로운 유형의 혁신을 자극할 수 있음

- AI의 개발, 배치 및 사용을 가장 잘 지원하는 방법을 이해하여 모든 사람이 AI 기반 세계에서 번창하고 더 나은 미래를 구축하는 동시에 글로벌 경쟁력을 확보하는 것이 중요함
- 다른 강력한 기술과 마찬가지로 우리 사회에서 AI 시스템을 사용하면 사람과 사회에 미치는 영향, 의사결정 능력 및 안전과 관련된 몇 가지 윤리적 문제가 발생함
- AI 시스템의 지원을 점점 더 많이 사용하거나 의사결정을 AI 시스템에 위임하려

는 경우, 그 시스템이 사람들의 삶에 미치는 영향이 공정한지, 위태롭지 않게 하고 그에 따라 행동할 수 있는 가치관에 부합하는지, 그리고 적절한 책임 프로세스가 이를 보장할 수 있는지를 확인할 필요가 있음
- 유럽은 실현하고자 하는 AI 미래의 규범적 비전을 정의하고, 이 비전을 달성하기 위해 유럽에서 어떤 AI 개념을 연구, 개발, 배포 및 사용해야 하는지 이해해야 할 필요가 있음
- 이 보고서를 통해 우리는 AI로 미래를 구축하는 올바른 방법이라고 믿는 신뢰할 수 있는 AI 개념을 도입함으로써 이러한 노력에 이바지하고자 함
- 민주주의, 법치 및 기본권이 AI 시스템을 뒷받침하고, 그러한 시스템이 지속적으로 민주적 문화를 옹호하고 개선하는 미래는 혁신과 책임있는 경쟁력이 번창할 수 있는 환경을 가능하게 할 것임
- 한정된 영역에 관한 윤리 법령은 일관되고 세분화된 버전이라도 상황별 세부 사항에 항상 민감해야만 하는 윤리적 추론 그 자체를 대체하는 것으로서 기능할 수 없음
- 일련의 규칙을 개발하는 것 외에도 신뢰할 수 있는 AI를 보장하려면 공개 토론, 교육 및 실제 학습을 통해 윤리적 문화와 사고방식을 구축하고 유지해야 함

1) 도덕적 · 법적 자격으로서의 기본권

- 우리는 유럽연합 조약, 유럽연합 헌장 및 국제 인권법에 명시된 기본권에 기반을 둔 AI 윤리에 대한 접근법을 신뢰함
- 민주주의와 법치의 체계 내에서 기본권에 대한 존중이 AI 환경에서 운영될 수 있는 추상적인 윤리적 원칙과 가치를 밝힐 수 있는 가장 유망한 방법을 제공함
- 유럽연합 조약 및 유럽연합 헌장은 유럽연합 법률을 시행할 때 유럽연합 회원국과 유럽연합 기관이 법적으로 존중해야 하는 일련의 기본 권리를 규정하며, 이러한 권리는 존엄성, 자유, 평등 및 연대, 시민의 권리 및 정의와 관련하여 유럽연합 헌장에 묘사되어있음
- 이러한 권리를 통합하는 공통의 토대는 인간 존엄성에 뿌리를 두고 있는 것으로 이해될 수 있으며, 이것에 의해 민간, 정치적, 경제적, 사회적 분야에서 인간이 유일하고 양도할 수 없는 도덕적 지위를 누린다고 하는 인간 중심적 접근으로서 우리가 설명하는 것을 나타낼 수 있음
- 유럽연합 헌장에 명시된 권리는 법적 구속력이 있지만, 기본권이 모든 경우에 포괄적인 법적 보호를 제공하지는 않고, 유럽연합 헌장의 경우 적용 분야가 유럽연합 법률 영역으로 제한됨
- 국제 인권법, 특히 유럽 인권 협약은 유럽연합 법률의 범위를 벗어나는 영역을 포함하여 유럽연합 회원국에 대해 법적 구속력을 가지며, 동시에 인간으로서의 도덕적 지위에 따라 법적 힘과는 별개로 개인과 집단에게 기본권이 부여됨

- 법적으로 집행 가능한 권리로 이해된다면, 기본권은 법률 준수를 보호하는 신뢰할 수 있는 AI(합법적 AI)의 첫 번째 구성요소에 속하지만, 인간의 고유한 도덕적 지위에 뿌리를 둔 모든 사람의 권리로 이해된다면, 그들은 또한 신뢰할 수 있는 AI(윤리적 AI)의 두 번째 구성요소를 뒷받침하며 법적 구속력은 없지만, 신뢰성을 보장하는 데 중요한 윤리적 규범이 됨
- 본 보고서는 전자 구성요소에 대한 가이드를 제공하는 것이 아니므로, 이러한 구속력이 없는 가이드의 목적을 위해 기본권에 대한 참조는 후자의 구성요소를 반영함

2) 기본권에서 윤리적 원칙까지

① 신뢰할 수 있는 AI에 대한 근거로서의 기본권

- 국제 인권법, 유럽연합 조약 및 유럽연합 헌장에 명시된 포괄적인 불가분 권리 중 아래의 기본권들은 AI 시스템을 다루는 데 특히 쉬움
- 특정 상황에서 이러한 권리들은 유럽연합에서 법적으로 집행 가능하므로, 해당 조건에 대한 준수는 법적으로 의무이지만, 법적으로 집행 가능한 기본권을 준수한 후에도, 윤리적 성찰은 AI 시스템의 개발, 배포 및 사용이 어떻게 기본권이나 근본적인 가치와 관련되는가를 보여줄 수 있으며, 우리가 현재 기술을 가지고 할 수 있는 것보다는 우리가 해야만 하는 것을 밝히려고 노력할 때 더 세분된 가이드를 제공하는데 도울 수 있음

<인간의 존엄성에 대한 존경>

- 인간의 존엄성은 모든 인간이 타인이나 AI 시스템과 같은 새로운 기술들에 의해 결코 감소, 타협, 억압되어서는 안 되는 '본질적 가치'를 가진다고 하는 생각을 포함하고 있음
- 이러한 환경에서, 인간 존엄성에 대한 존중은 모든 사람이 단순히 조사되거나, 분류되거나, 점수 매겨지거나, 몰려지거나, 길들여지거나, 조작의 대상이 되는 것이 아니라 도덕적 주체로 존중하여 취급되는 것을 수반함
- AI 시스템은 인간의 신체적·정신적 무결성, 개인 및 문화적 정체성 감각, 필수 요구사항의 만족을 존중하고, 봉사하며, 보호하는 방식으로 개발되어야 함

<개인의 자유>

- 인간은 스스로 인생의 결정을 내릴 자유를 가져야 하며, 이를 위해 주권적 침입으로부터의 자유가 필요하지만, 배제 위험에 처한 개인이나 사람들이 AI의 혜택과 기회에 동등하게 접근할 수 있도록 정부 및 비정부 조직의 개입도 요구됨
- 예를 들어, AI 환경에서 개인의 자유는 직접 또는 간접적 불법 강압, 정신적 자

율성과 정신적 건강에 대한 위협, 부당한 감시, 속임수 그리고 부당한 조작의 완화를 요구함

- 개인의 자유는 개인이 자신의 삶에 대해 더 높은 수준의 통제권을 행사할 수 있도록 하는 약속을 의미하며, 여기에는 비즈니스 수행의 자유, 예술과 과학의 자유, 표현의 자유, 개인적 생활에 대한 권리와 프라이버시, 집회 및 결사의 자유를 포함함

<민주주의, 공정성 그리고 법치에 대한 존중>

- AI 시스템은 민주적 프로세스를 유지 및 육성하고 개인의 다양한 가치와 삶의 선택을 존중해야 함
- AI 시스템은 민주적 프로세스, 인간의 심의 또는 민주적 투표 시스템을 약화시켜서는 안 됨
- AI 시스템은 법치의 기반이 되는 기초적인 약속, 필수 법률 및 규정을 훼손하는 방식으로 작동하지 않도록 하고 법 앞에서 적법한 절차와 평등을 보장하기 위한 약속을 포함해야 함

<평등, 비차별 그리고 연대 - 배제 위험이 있는 인간의 권리를 포함>

- 모든 인간의 도덕적 가치와 존엄성에 대한 동등한 존중이 보장되어야 함
- 이것은 객관적 정당성에 근거하여 서로 다른 상황을 구별하는 것을 허용하는 비차별을 넘어섬
- AI 환경에서 평등은 시스템 운영이 부당하게 편향된 출력을 생성할 수 없음을 의미하며, 예로써 AI 시스템을 훈련하는 데 사용된 데이터는 가능한 한 포괄적이어야 하며 다른 인구 그룹을 나타내야 함
- 근로자, 여성, 장애인, 소수 민족, 아동, 소비자 또는 배제 위험에 처한 다른 사람들과 같이 잠재적으로 취약한 개인 및 그룹에 대한 적절한 존중을 요구함

<시민의 권리>

- 시민들은 투표권, 행정권 또는 공공 문서에 대한 접근권, 행정부에 청원 할 권리를 포함한 다양한 권리로부터 혜택을 받음
- AI 시스템은 공공재와 서비스를 사회에 제공함에 있어 정부의 규모와 효율성을 개선할 수 있는 상당한 잠재력을 제공하지만, 시민의 권리는 AI 시스템에 의해 부정적인 영향을 받을 수 있음
- 여기에서 “시민권“이라는 용어가 사용된 경우, 이는 국제법에 따라 그래서 AI 시스템의 영역에서 권리를 보유한 EU 내 제3국 국민 및 불법적인 사람들의 권리를 거부하거나 무시하는 것이 아님

② AI 시스템의 환경에서 윤리적 원칙

- 많은 공공, 개인, 민간 조직이 AI 시스템에 대한 윤리적 체계를 생성하기 위해 기본권에서 영감을 얻어왔음
- 유럽연합 내에서, 과학 및 신기술 윤리에 관한 유럽 그룹(European Group on Ethics in Science and New Technologies, "EGE")은 EU 조약 및 헌장에 규정된 기본적인 가치에 바탕을 둔 9가지 기본원칙을 제안했음
- 우리는 이 작업을 더욱 발전시켜 다양한 그룹이 지금까지 제시한 대부분 원칙을 인정하고 모든 원칙이 육성하고 지원하려는 목적을 명확히 함
- 이러한 윤리적 원칙은 새롭고 구체적인 규제 도구에 영감을 줄 수 있으며, 시간이 지남에 따라 사회 기술 환경이 진화할 때 기본 권리를 해석하는 데 도움이 될 수 있으며, AI 시스템의 개발, 배포 및 사용에 대한 근거를 제공할 수 있으며 사회 자체가 진화함에 따라 동적으로 적응할 수 있음
- AI 시스템은 개인 및 집단 복지를 개선해야 함
- 본 장에서는 AI 시스템이 신뢰할 수 있는 방식으로 개발, 배포 및 사용되도록 보장하기 위해 존중되어야 하는 기본권에 뿌리를 둔 네 가지 윤리 원칙이 나열되어 있음
- AI 실무자가 항상 준수하기 위해 노력해야 하는 윤리적 명령으로 지정되며, 이에는
  ⅰ. 인간의 자율성에 대한 존중
  ⅱ. 피해 방지
  ⅲ. 공정성
  ⅳ. 설명 가능성 원칙들임
- 이들 중 상당수는 의무적 준수가 요구되는 기존 법적 요건에 이미 상당 부분 반영되어 있으므로 신뢰할 수 있는 AI의 첫 번째 구성요소인 합법적 AI의 범위에 포함됨
- 그러나 위에서 설명한 대로 많은 법적 의무 윤리 원칙을 반영하고 윤리 원칙을 준수하는 것은 기존 법률을 공식적으로 준수하는 것 이상임

<인간의 자율성에 대한 존중 원칙>

- EU가 기초하고 있는 기본권은 인간의 자유와 자율성에 대한 존중을 보장하는 데 있음
- AI 시스템과 상호 작용하는 인간은 자신에 대해 완전하고 효과적인 자기 결정을 유지할 수 있어야 하며 민주적 과정에 참여할 수 있어야 함
- AI 시스템은 인간을 부당하게 종속시키거나, 강요하거나, 속이거나, 조작하거나, 조절하거나, 몰아서는 안 되며, 인간의 인지적, 사회적, 문화적 기술을 증강, 보완 및 강화하도록 설계되어야 함
- 인간과 AI 시스템 간의 기능 할당은 인간 중심의 설계 원칙을 따라야 하며 인간

의 선택을 위한 의미 있는 기회를 남겨야 하고, 이는 AI 시스템의 작업 프로세스에 대한 인간의 감독을 확보하는 것을 의미함
- AI 시스템은 근본적으로 작업 영역을 바꿀 수 있고, 작업 환경에서 인간을 지원하고 의미 있는 작업 창출을 목표로 해야 함

<피해 방지의 원칙>
- AI 시스템은 피해를 유발하거나 악화시켜서는 안 되며, 인간에게 불리한 영향을 주어서는 안 됨
- 이는 인간의 존엄성과 정신적·육체적 무결성을 보호하는 것을 수반함
- AI 시스템과 운영 환경은 안전하고 확실해야 하며, 기술적으로 견고해야 하며 악의적인 사용에 개방되지 않도록 해야 함
- 취약한 사람들이 더 많은 관심을 받고, AI 시스템의 개발, 배포 및 사용에 포함되어야 함
- AI 시스템이 고용주와 직원, 기업과 소비자 또는 정부와 시민 간의 힘이나 정보의 비대칭으로 인해 악영향을 유발하거나 악화시킬 수 있는 상황에도 특히 주의해야 함
- 피해를 방지하려면 자연환경과 모든 생명체를 고려해야 함

<공정성의 원칙>
- AI 시스템의 개발, 배치 및 사용은 공정해야 함
- 공정성은 실질적이고 절차적인 차원을 모두 가지고 있음
- 실질적인 차원은 혜택과 비용의 동등하고 공정한 분배를 보장하고, 개인과 그룹이 불공정한 편견, 차별 및 낙인으로부터 자유로워지도록 보장하기 위한 약속을 포함함
- 공정성은 AI 실무자가 수단과 목적 사이의 비례 원칙을 존중하고, 경쟁하는 이익과 목표의 균형을 맞추는 방법을 신중하게 고려해야 함을 의미함
- 공정성의 절차적 차원은 AI 시스템과 인간이 그것들을 운영함으로써 내려진 결정에 대해 경쟁하고 효과적인 보상을 추구할 수 있는 능력을 수반함
- 이를 위해, 결정에 대한 책임 있는 주체가 식별되어야만 하고 의사결정 과정이 설명될 수 있어야 함

<설명 가능성의 원칙>
- 설명 가능성은 AI 시스템에 대한 사용자의 신뢰를 구축하고 유지하는 데 중요함
- 이는 프로세스가 투명해야 하고, AI 시스템의 역량과 목적이 공개적으로 전달되어야 하며, 가능한 한 직간접적으로 영향을 받는 사람들에게 설명 가능한 결정이 필요함을 의미함
- 그러한 정보가 없다면, 결정은 적절한 절차에 따라 정당하게 다투어질 수 없음

- 모델이 특정 결과 또는 결정을 생성한 이유가 항상 가능한 것은 아니며, 이러한 경우를 '블랙박스' 알고리즘이라고 하며 특별한 주의가 필요함
- 이러한 상황에서 시스템 전체가 기본적 권리를 존중한다면, 다른 설명 조치(예: 추적성, 감사 가능성 및 시스템 역량에 대한 투명한 의사소통)가 필요할 수 있으며, 설명이 필요한 정도는 출력이 잘못되었거나 부정확한 경우 결과의 심각성과 환경에 따라 크게 달라짐

③ 원칙들 사이의 충돌

- 확립된 솔루션이 없는 위의 원칙 사이에 충돌이 발생할 수 있으며, 민주적 참여, 적법 절차 및 개방적인 정치 참여에 대한 EU의 근본적인 기여에 따라 이러한 충돌에 대처하기 위한 책임 있는 숙고 방법이 확립되어야 함
- 예를 들어 다양한 응용 분야에서, 피해 방지 원칙과 인간 자율성의 원칙이 상충될 수 있음: 예를 들어, 범죄를 줄이는 데 도움이 될 수 있지만, 개인의 자유와 프라이버시를 침해하는 감시 활동을 수반하는 방식으로 '예측 치안'의 AI 시스템을 사용하는 경우
- 더 나아가, AI 시스템의 전반적인 혜택은 예측 가능한 개별 위험을 많이 초과해야 함
- 위의 원칙은 확실히 해결책에 대한 지침을 제공하지만, 추상적인 윤리적 처방으로 남아 있어서, AI 실무자는 위의 원칙들에 바탕을 둔 올바른 솔루션을 찾을 것이라고 기대될 수 없지만, 그들은 직관이나 무작위 재량보다는 합리적이고 증거에 기반을 둔 반성을 통해 윤리적 딜레마와 절충점에 접근해야 함

<1장의 주요 가이드>

- 윤리적 원칙(인간의 자율성 존중, 피해 방지, 공정, 설명가능)을 준수하는 방법으로 AI 시스템을 개발, 배치, 이용하라
- 원칙들 사이의 잠재적 충돌을 인정하고 해결하라
- 어린이, 장애인 그리고 역사적으로 불이익을 받아오거나 배제될 위험에 처해 있는 사람들과 같이 더 취약한 그룹과 관련된 상황 그리고 고용주와 종업원이나 비즈니스와 소비자들 사이와 같이 권력과 정보의 불균형으로 특정화되는 상황들에 각별한 주의를 기울여라
- 개인과 사회에 상당한 혜택을 가져올지라도, AI 시스템이 특정 위험을 야기하고, 예상하거나, 밝히거나 측정하기 어려울 수도 있는 영향(즉 민주주의, 법의 지배와 분배의 공정 또는 인간의 마음 그 자체에 대한)을 포함하여 부정적인 영향을 가져올 수 있다는 것을 인정하라

- 적절할 때 그리고 그 위험의 크기와 비례하여 이러한 위험들을 경감시키는 적절한 조치들을 채택하라

(3) 신뢰할 수 있는 AI의 실현

- 본 장은 1장에서 요약된 원칙을 기반으로 충족되어야 하는 7가지 요구사항 목록을 통해 신뢰할 수 있는 AI의 구현 및 실현에 대한 지침을 제공함
- AI 시스템의 수명주기 전반에 걸쳐 이러한 요구사항들의 실행을 위해 이용 가능한 기술적·비기술적 방법들도 소개됨

1) 신뢰할 수 있는 AI의 요건

- 1장에 설명된 원칙은 신뢰할 수 있는 AI를 달성하기 위한 구체적인 요구사항이며, 이러한 요구사항은 AI 시스템의 수명주기에 참여하는 다양한 이해 관계자(개발자, 배포자 및 최종 사용자는 물론 더 넓은 사회)에게 적용됨[2)]
- 이해 관계자 그룹마다 요구사항이 충족되도록 하는 역할이 다름: ⅰ. 개발자는 절차를 설계하고 개발하는 요건들을 실행하고 적용해야 함; ⅱ. 배포자는 사용하는 시스템과 제공하는 제품 및 서비스가 요구사항을 충족하는지 확인해야 함; ⅲ. 최종 사용자와 더 넓은 사회는 이러한 요구사항에 대한 정보를 얻고 이를 유지하도록 요청할 수 있어야 함
- 아래의 요구사항 목록은 완전하지 않지만, 체계적, 개별적 및 사회적 측면이 포함됨
  - ⅰ. 인간의 개입과 감독: 기본권, 인간의 개입 및 감독 포함
  - ⅱ. 기술적 견고성과 안전성: 공격 및 보안에 대한 복원력, 만일에 대한 계획 및 일반 안전, 정확성, 신뢰성 및 재현성 포함
  - ⅲ. 개인 정보 보호 및 데이터 거버넌스: 프라이버시 존중, 데이터의 품질 및 무결성, 데이터 접근 포함
  - ⅳ. 투명성: 추적성, 설명성 및 커뮤니케이션 포함
  - ⅴ. 다양성, 차별 금지 및 공정성: 불공정한 편견 방지, 접근성 및 유니버설 디자인, 이해 관계자 참여 포함
  - ⅵ. 사회적·환경적 복지: 지속 가능성 및 환경 친화성, 사회적 영향, 사회 및 민주주의 포함
  - ⅶ. 책무: 감사 가능성, 부정적인 영향 최소화 및 보고, 균형 및 보상을 포함

---

2) 개발자는 AI 시스템을 연구, 설계 또는 개발하는 사람들을 말하고, 배포자는 비즈니스 프로세스 내에서 AI 시스템을 사용하고 다른 사람에게 제품과 서비스를 제공하는 공공 또는 민간 조직을 말하며, 최종 사용자는 직접 또는 간접적으로 AI 시스템에 참여하는 사람들이며, 더 넓은 사회는 AI 시스템에 의해 직간접적으로 영향을 받는 다른 모든 사회를 포괄한다.

- 모든 요구사항이 똑같이 중요하지만, 서로 다른 영역과 산업에 적용될 때 상황과 잠재적인 충돌을 고려해야 함
- 이러한 요구사항의 구현은 AI 시스템의 전체 수명주기 동안 발생해야 하며, 구체적인 적용에 따라 다름
- 대부분 요구사항이 모든 AI 시스템에 적용되지만, 직간접적으로 영향을 미치는 사람들에게 특별한 주의가 주어지며, 일부 응용 프로그램의 경우 관련성이 떨어질 수 있음
- 위의 요구사항은 때에 따라 기존 법률에 이미 반영된 요소가 포함되며, 신뢰할 수 있는 AI의 첫 번째 구성요소에 따라 수평적으로 적용되는 규칙과 영역별 규정에 관한 법적 의무를 준수하는 것이 AI 실무자의 책임임

① 인간의 개입과 감독

- 인간의 자율성 존중의 원칙에 의해 설명된 바와 같이, AI 시스템은 인간의 자율성과 의사결정을 지지해야 하며, 이는 AI 시스템이 사용자의 개입을 지지함으로써 민주적이고, 번영되며, 평등한 사회에 대한 가능자로서 행동하고 기본권을 신장시켜야 하며, 인간의 감독을 허용하는 것을 요구함

<기본권>

- 많은 기술과 마찬가지로 AI 시스템은 기본권을 활성화할 수도 방해할 수도 있음
- 예를 들어 개인 데이터를 추적하도록 돕거나 교육 접근성을 높여 교육에 대한 권리를 지원함으로써 사람들에게 혜택을 줄 수 있지만, AI 시스템의 범위와 용량을 고려할 때, 기본권에 부정적인 영향을 미칠 수도 있음
- 그러한 위험이 존재하는 상황에서는 기본권 영향 평가를 수행해야 하며, 이는 시스템의 개발에 앞서 수행되어야 하고, 그 위험이 다른 사람의 권리와 자유를 존중하기 위하여 민주 사회에서 필요에 따라 감소 또는 정당화될 수 있는지에 대한 평가를 포함해야 함
- 메커니즘은 잠재적으로 기본권을 침해하는 AI 시스템에 대한 외적 피드백을 수신할 장소를 마련해야만 함

<인간의 개입>

- 사용자는 AI 시스템에 대하여 정보를 바탕으로 자율적 결정을 내릴 수 있어야 하며, AI 시스템을 충분히 이해하고 상호 작용할 수 있는 지식과 도구를 받아야 하며, 가능한 경우 시스템을 자체 평가하거나 이의를 제기할 수 있어야 함
- AI 시스템은 목표에 따라 더 나은, 더 많은 정보에 따른 선택을 할 수 있도록 개인을 지원해야 함
- AI 시스템은 개인의 자율성을 위협할 수 있는 다양한 형태의 불공정한 조작, 속

임수, 몰기 및 훈련을 포함한 잠재의식 프로세스를 이용할 수 있어, 감지하기 어려울 수 있는 메커니즘을 통해 인간 행동을 형성하고 영향을 미치기 위해 때로 배치될 수 있음

- 사용자 자율성의 전반적인 원칙은 시스템 기능의 중심이 되어야 하며, 이것의 핵심은 사용자에게 법적 영향을 미치거나 유사하게 그들에게 상당한 영향을 미치는 경우 자동화된 처리만을 기반으로 한 결정을 받지 않을 권리임

<인간의 감독>

- 인간의 감독은 AI 시스템이 인간의 자율성을 훼손하거나 기타 부작용이 발생하지 않음을 보장하는 데 도움이 됨
- 인감참여(human-in-the-loop, HITL), 인간지배(human-on-the loop, HOTL), 또는 인간지휘(human-in-command, HIC) 접근과 같은 거버넌스 메커니즘을 통해 감독은 성취될 수 있음
- HITL은 시스템의 모든 의사 결정주기에서 사람이 개입할 수 있는 능력을 의미하며, 대부분의 경우 가능하지도 바람직하지도 않음
- HOTL은 시스템 설계 주기 동안 사람이 개입하고 시스템 작동을 관찰할 수 있는 능력을 나타냄
- HIC는 AI 시스템의 전반적인 활동 (광범위한 경제적, 사회적, 법적 및 윤리적 영향 포함)을 감독하는 능력과 특정 상황에서 시스템을 언제 어떻게 사용할지 결정할 수 있는 능력을 나타냄
- 여기에는 특정 상황에서 AI 시스템을 사용하지 않기로 한 결정, 시스템 사용 중에 인간의 재량 수준을 설정하거나 시스템에서 내린 결정을 무시할 수 있는 능력을 보장하는 결정이 포함될 수 있으며, 공무원이 자신의 임무에 따라 감독할 수 있는 능력을 갖는 것을 보장해야 함
- AI 시스템의 응용 분야 및 잠재적 위험에 따라 안전 및 제어 조치를 지원하기 위해 다양한 수준의 감독 메커니즘이 필요할 수 있으며, 인간이 AI 시스템에 대해 실행할 수 있는 감독이 적을수록 더 광범위한 테스트와 더 엄격한 거버넌스가 필요함

② 기술적 견고성과 안전성

- 신뢰할 수 있는 AI를 달성하는 데 중요한 요소는 기술적 견고성이며, 이는 피해 방지 원칙과 밀접하게 관련되어 있음
- 기술적 견고성을 위해서는 AI 시스템이 위험에 대한 예방적 접근 방식과 의도한 대로 안정적으로 작동하는 동시에 의도하지 않은 예상치 못한 피해를 최소화하고 허용할 수 없는 피해를 방지하는 방식으로 개발되어야 함
- 이것은 또한 운영 환경의 변화나 적대적인 방식으로 시스템에 작용할 수 있는

다른 기관의 존재에 대한 잠재적인 변화에도 적용되어야 하며, 인간의 육체적·정신적 완전성이 보장되어야 함

<공격에 대한 탄력성 및 보안>

- 모든 소프트웨어 시스템과 마찬가지로 AI 시스템은 해킹과 같은 공격자가 악용할 수 있는 취약성으로부터 보호되어야 함
- 공격은 데이터(data poisoning), 모델(model leakage) 또는 기본 인프라(SW와 HW)를 표적으로 삼을 수 있음
- AI 시스템이 공격을 받는 경우 즉, 적대적 공격의 경우, 데이터 및 시스템 동작이 변경되어 시스템이 다른 결정을 내리거나 완전히 종료될 수 있음
- 시스템 및 데이터는 악의적인 의도에 의해 예상치 못한 상황에 노출되어 손상될 수 있으며, 불충분한 보안 프로세스는 잘못된 결정 또는 물리적 피해 초래할 수 있음
- AI 시스템이 안전한 것으로 간주 되려면, AI 시스템의 의도하지 않은 적용과 악의적 행위자에 의한 시스템의 잠재적 남용이 고려되어야만 하며, 이들을 막고 완화하는 조치들이 취해져야 함

<대체 계획과 일반적인 안전>

- AI 시스템은 문제가 있는 경우에 대체 계획을 가능케 하는 안전장치를 가져야만 하며, 이는 AI 시스템이 통계에서 규칙 기반 절차로 전환하거나 행동을 계속하기 전에 인간 운영자를 요청한다는 것을 의미할 수 있음
- 시스템은 생명체나 환경을 해치지 않고 해야 할 일을 할 수 있도록 보장해야 하며, 이는 의도하지 않은 결과 및 오류의 최소화를 포함함
- 다양한 응용 분야에서 AI 시스템 사용과 관련된 잠재적 위험을 명확히 하고 평가하는 프로세스를 수립해야 함
- 필요한 안전 조치의 수준은 AI 시스템이 제기하는 위험의 규모에 따라 달라지며, 이는 시스템의 능력에 따라 달라지며, 개발 프로세스 또는 시스템 자체가 특히 높은 위험을 초래할 것으로 예상하는 경우 안전 조치를 사전에 개발하고 테스트하는 것이 중요함

<정확성>

- 정확성은 올바른 판단을 내리는 AI 시스템의 능력(예: 정보를 적절한 카테고리로 올바르게 분류하는 능력, 데이터 또는 모델을 기반으로 올바른 예측, 권장 사항 또는 결정을 내리는 능력)과 관련이 있음
- 명확하고 잘 구성된 개발 및 평가 프로세스는 부정확한 예측으로 인한 의도하지 않은 위험을 완화 및 수정할 수 있음
- 가끔 부정확한 예측을 피할 수 없는 경우, 시스템이 이러한 오류의 가능성을 나

타낼 수 있는 것이 중요함
- AI 시스템이 인간의 삶에 직접적인 영향을 미치는 상황에서는 높은 수준의 정확성이 특히 중요함

<신뢰성 및 재현성>
- AI 시스템의 결과가 재현 가능할 뿐만 아니라 믿을 수 있는 것이 중요함
- 믿을 수 있는 AI 시스템은 다양한 입력과 다양한 상황에서 제대로 작동하는 시스템이고, 이는 AI 시스템을 면밀히 검토하고 의도하지 않은 피해를 방지하는 데 필요함
- 재현성은 같은 조건으로 반복될 때 AI 실험이 같은 행동을 보이는지에 관한 것이며, 이것은 과학자들과 정책 입안자들로 하여금 AI 시스템이 할 수 있는 것을 정확하게 설명할 수 있게 함
- 복제 파일은 동작을 테스트하고 재현하는 프로세스를 용이하게 할 수 있음

③ 프라이버시와 데이터 거버넌스

- AI 시스템에 의해 특히 영향 받는 기본권인 프라이버시는 피해 방지의 원칙과 면밀하게 관련되어 있음
- 프라이버시에 대한 피해의 방지는 사용된 데이터의 품질과 무결성을 다루는 적절한 데이터 거버넌스, AI 시스템이 배치될 영역 측면에서 적절성, 접속 규약 그리고 프라이버시를 보호할 방법으로 데이터를 처리하는 능력이 필요함

<프라이버시 및 데이터 보호>
- AI 시스템은 전체 수명주기 동안 프라이버시와 데이터 보호를 보장해야 함
- 여기에는 사용자가 처음에 제공한 정보와 시스템과의 상호 작용 과정에서 사용자에 대해 생성된 정보가 포함됨
- 인간 행동에 대한 디지털 기록을 통해 AI 시스템은 개인의 선호도뿐만 아니라 성적 취향, 연령, 성별, 종교 또는 정치적 견해를 추론할 수 있음
- 개인들이 데이터 수집 프로세스를 신뢰할 수 있도록 허용하기 위해, 수집된 데이터는 불법적이거나 부당하게 차별하는 데 사용되지 않도록 해야 함

<품질과 데이터의 무결성>
- 사용되는 데이터 세트의 품질은 AI 시스템의 성능에 가장 중요함
- 데이터가 수집될 때, 사회적으로 구성된 편견, 부정확성, 오류 및 실수가 포함될 수 있으며, 이는 주어진 데이터 세트로 훈련하기 전에 해결해야 함
- 또한 데이터의 무결성이 보장되어야 함
- 악성 데이터를 AI 시스템에 공급하면 특히 자가 학습 시스템에서 행동이 바뀔

수 있으므로, 사용되는 프로세스와 데이터 세트는 기획, 교육, 테스트 및 배치와 같이 각 단계에서 테스트 되고 기록되어야만 함
- 이것은 또한 자체 개발한 것이 아니라 다른 곳에서 획득된 AI 시스템에도 적용되어야 함

<데이터에 대한 접근>
- 개인의 데이터를 처리하는 조직(누가 시스템 사용자인지 여부에 관계없이)에는 데이터 접근을 관리하는 데이터 규약이 마련되어 있어야 함
- 이러한 규약은 누가 어떤 상황에서 데이터에 접근할 수 있는지를 설명해야 함
- 역량을 갖고 개인의 데이터에 접근할 필요가 있는 정식자격이 있는 직원만이 그렇게 하도록 허락되어야만 함

④ 투명성

- 이 요건은 설명 가능성의 원칙과 면밀하게 관련되며, AI 시스템과 관련된 요소들(데이터, 시스템, 비즈니스 모델)의 투명성을 포함함

<추적성>
- 사용되는 알고리즘뿐만 아니라 데이터 수집과 데이터 라벨 표시를 포함하여, AI 시스템의 결정을 만들어내는 데이터 세트와 프로세스는 추적성과 투명성의 증가를 허용하는 가능한 한 최고의 기준으로 작성되어야 함
- 이것은 또한 AI 시스템에 의한 결정에 적용되며, 이것은 AI 결정이 왜 잘못되었고, 그 후의 실수를 막는 데 도움을 줄 수 있는 이유를 식별할 수 있게 함
- 추적 가능성은 감사 가능성과 설명 가능성을 촉진함

<설명 가능성>
- 설명 가능성은 AI 시스템의 기술적 프로세스와 관련된 인간의 결정 둘 다를 설명하는 능력과 관련이 있음
- 기술적 설명 가능성은 AI 시스템이 내린 결정을 인간이 이해하고 추적할 수 있어야 함
- 시스템의 설명 가능성을 높이거나(정확도를 낮출 수 있음) 정확도를 높이는 것(설명 가능성을 희생) 사이에 절충안을 만들어야 할 수도 있음
- AI 시스템이 사람들의 삶에 중대한 영향을 미칠 때마다 AI 시스템의 의사결정 과정에 대한 적절한 설명을 요구할 수 있어야 함
- 그러한 설명은 시의적절해야 하며 관련 이해 관계자(예: 비전문가, 규제 기관 또는 연구원)의 전문 지식에 맞게 조정되어야 함
- AI 시스템이 조직의 의사결정 프로세스, 시스템의 설계 선택 및 배포 이유에 영

향을 미치고 형성하는 정도에 대한 설명이 이용 가능하여야 하며, 이에 따라 비즈니스 모델 투명성을 보장해야 함

<커뮤니케이션>

- AI 시스템은 사용자에게 자신을 인간으로 나타내서는 안 되며, 인간은 AI 시스템과 상호 작용하고 있다는 것을 통보받을 권리를 갖고 있음
- 이는 AI 시스템이 확인 가능해야만 한다는 것이며, 기본권 준수를 보장할 필요가 있는 경우 인간 상호 작용을 위해 이러한 상호 작용을 반대하는 옵션이 제공되어야 함
- 이 외에도 AI 시스템의 기능과 한계는 당면한 사용 사례에 적합한 방식으로 AI 실무자 또는 최종 사용자에게 전달되어야 하며, 여기에는 AI 시스템의 정확도 수준과 한계에 대한 커뮤니케이션이 포함될 수 있음

⑤ 다양성, 비차별 그리고 공정성

- 신뢰할 수 있는 AI를 달성하려면 전체 AI 시스템의 수명주기 전반에 걸쳐 포용성과 다양성을 구현해야 합니다. 프로세스 전반에 걸쳐 영향을 받는 모든 이해관계자의 고려와 참여 외에도 포괄적인 설계 프로세스와 동등한 대우를 통해 동등한 접근을 보장합니다. 이 요구 사항은 공정성의 원칙과 밀접한 관련이 있습니다.

<불공정한 편견의 회피>

- AI 시스템에서 사용하는 데이터 세트(교육 및 운영 모두)는 의도하지 않은 역사적 편견, 불완전성 및 잘못된 거버넌스 모델이 포함되어 어려움을 겪을 수 있음
- 이러한 편견의 지속은 특정 집단이나 사람들에 대한 의도하지 않은 (간접) 편견과 차별로 이어질 수 있으며, 편견과 소외를 잠재적으로 악화시킬 수 있음
- 피해는 (소비자) 편견의 의도적인 착취 또는 공모나 불투명한 시장의 방법에 의한 가격의 균질화와 같은 불공정 경쟁에 의해 발생할 수 있음
- 식별 가능하고 차별적인 편견은 가능한 한 수집 단계에서 제거되어야 하며, AI 시스템이 개발되는 방식(예: 알고리즘의 프로그래밍)도 불공정한 편견을 겪을 수 있음
- 이는 시스템의 목적, 제약, 요구 사항 및 결정을 명확하고 투명한 방식으로 분석하고 해결하기 위한 감독 프로세스를 배치함으로써 대응할 수 있으며, 다양한 배경, 문화 및 규율의 채용은 의견의 다양성을 보장할 수 있으며 장려되어야 함

<접근성 및 보편적인 디자인>

- 특히 B2C 영역에서 시스템은 사용자 중심이어야 하며 연령, 성별, 능력 또는 특

성과 관계없이 모든 사람이 AI 제품 또는 서비스를 사용할 수 있도록 설계되어야 함

- 모든 사회 집단에 존재하는 장애인을 위한 기술의 접근성은 특히 중요함
- AI 시스템은 일률적 접근 방식을 가져서는 안 되며 관련 접근성 표준에 따라 가능한 가장 광범위한 사용자를 다루는 보편적인 디자인 원칙을 고려해야 함
- 이는 동등한 접근과 기존과 나타나는 컴퓨터 매개 인간 활동과 보조기술과 관련하여 모든 사람의 적극적인 참여를 가능케 할 것임

<이해 관계자 참여>

- 신뢰할 수 있는 AI 시스템을 개발하려면 수명주기 동안 시스템에 직간접적으로 영향을 받을 수 있는 이해 관계자와 컨설트하는 것이 필요함
- 배치 후에도 정기적인 피드백을 요청하고 조직에서 AI 시스템을 구현하는 전체 프로세스에 걸쳐 근로자 정보, 상담 및 참여를 보장함으로써 이해 관계자 참여를 위한 장기적인 메커니즘을 설정하는 것이 필요함

⑥ 사회적 그리고 환경적 복지

- 공정성과 피해 방지 원칙에 따라, 더 넓은 사회, 지각 있는 다른 존재들과 환경도 AI 시스템의 수명주기 전반에 걸쳐 이해 관계자로 간주되어야 함
- AI 시스템의 지속 가능성과 생태학적 책임은 장려되어야 하고, 지속 가능한 개발 목표와 같은 글로벌 관심 영역을 다루는 AI 솔루션에 관한 연구가 촉진되어야 함
- 이상적으로는 AI 시스템을 사용하여 미래 세대를 포함한 모든 인간에게 혜택을 주어야 함

<지속 가능하고 환경 친화적인 AI>

- AI 시스템은 가장 시급한 사회적 문제를 해결하는 데 도움이 될 것이지만, 이는 가장 환경친화적인 방식으로 이루어질 것을 보장해야 함
- 전체 공급망을 포함하여 시스템의 개발, 배포 및 사용 프로세스는 이러한 측면 즉, 자원 사용의 비판적인 검사와 훈련 동안 에너지 소비, 덜 해로운 선택의 결정을 통해 평가되어야만 함
- AI 시스템 전체 공급망의 환경 친화성을 확보하는 조치가 권장되어야 함

<사회적 영향>

- 우리 삶의 모든 영역(교육, 연구, 진료 또는 엔터테인먼트)에서 사회적 AI 시스템에 대한 보편적인 노출은 사회 기관에 대한 개념을 바꾸거나 사회적 관계 및 믿음에 영향을 미칠 수 있음

- AI 시스템은 사회적 기술을 향상하는 데도 사용될 수 있지만, 악화에도 똑같이 이바지할 수 있어, 사람들의 신체적·정신적 복지에 영향을 미칠 수 있음
- 이러한 시스템의 효과는 따라서 주의 깊게 모니터링하고 고려되어야 함

<사회와 민주주의>
- 개인에 대한 AI 시스템의 개발, 배포 및 사용의 영향을 평가하는 것 외에도, 이러한 영향은 기관, 민주주의 그리고 사회 전반에 대한 영향을 고려하면서 사회적 관점에서 평가되어야만 함
- AI 시스템의 사용은 정치적 의사결정뿐만 아니라 선거 상황도 포함하여, 특히 민주적 절차에 관한 상황에서 신중하게 고려되어야 함

⑦ 책무

- 책무의 요구사항은 위의 요구사항을 보완하고, 공정성의 원칙에 밀접하게 연결되어 있음
- 이 메커니즘은 개발, 배포 및 사용 전과 후 모두에 AI 시스템과 그 결과에 대한 책임과 책무를 보장하기 위해 시행되는 것이 필요함

<감사 가능성>
- 감사 가능성은 알고리즘, 데이터 및 설계 프로세스의 평가에 대한 가능성을 포함함
- 이것은 비즈니스 모델에 대한 정보와 AI 시스템과 관련된 지식재산이 항상 공개적으로 이용 가능해야만 한다는 것을 필연적으로 시사하지는 않음
- 내부 및 외부 감사자의 평가와 이러한 평가 보고서의 이용 가능성은 기술의 신뢰성에 이바지할 수 있음
- 안전에 중요한 애플리케이션을 포함하여 기본권에 영향을 미치는 응용 프로그램에 대해서는 AI 시스템이 독립적으로 감사 할 수 있어야 함

<부정적인 영향의 최소화 및 보고>
- 특정 시스템 결과에 이바지하는 조치 또는 결정에 대해 보고하는 능력과 그러한 결과에 대응하는 능력이 모두 보장되어야 함
- AI 시스템의 잠재적인 부정적 영향을 식별, 평가, 문서화 및 최소화하는 것은 (간접)직접적으로 영향을 받는 사람들에게 특히 중요함
- AI 시스템에 대한 합법적인 우려를 보고 할 때, 내부 고발자, NGO, 노동조합 또는 기타 단체에 대해 적절한 보호를 제공해야 함
- AI 시스템의 개발, 배포 및 사용 전과 도중의 영향 평가에 대한 사용은 부정적인 영향을 최소화하는 데 도움이 될 수 있으며, AI 시스템이 제기하는 위험에

비례해야 함

<균형(trade-off)>

- 위의 요구사항을 구현할 때, 둘 사이에 충돌이 발생할 수 있으며, 이는 불가피한 절충을 초래할 수 있음
- 이러한 균형은 그 시점에서의 기술적 수준 내에서 합리적이고 방법론적인 방식으로 해결되어야 하고, AI 시스템과 관련된 관련 이익과 가치들이 확인되어야 하며, 만일 충돌이 발생하는 경우, 기본권을 포함하여 윤리적 원칙에 대한 위험 측면에서 균형이 외적으로 인식되고 평가되는 것을 수반함
- 윤리적으로 허용되는 절충안을 식별할 수 없는 상황에서는 AI 시스템의 개발, 배치 및 사용이 그러한 형태로 진행되어서는 안 되며, 어떤 균형이 이루어져야 하는지에 대한 결정은 추론되고 적절하게 기록되어야 함
- 의사 결정자는 적절한 균형이 이루어지고 있는 방식에 대한 책임을 져야 하며, 필요한 변화가 필요한 시스템에서 이루어질 수 있는 것을 보장하는 결과적 결정의 적절성을 지속적으로 검토해야만 함

<보상>

- 부당한 악영향이 발생하면 적절한 보상을 보장하는 메커니즘이 예상되어야 함
- 일이 잘못되었을 때 보상이 가능하다는 것을 아는 것은 신뢰를 보장하기 위해 중요함
- 취약한 개인이나 그룹에 특별한 주의가 이루어져야 함

2) 신뢰할 수 있는 AI를 실현하기 위한 기술적·비기술적 방법

- 위의 요구사항을 구현하기 위해 기술적 방법과 비기술적 방법이 모두 사용될 수 있으며, 여기에는 AI 시스템 수명주기의 모든 단계가 포함됨
- 요구사항을 구현하는 데 사용되는 방법에 대한 평가와 구현 프로세스의 변경 사항을 보고하고 정당화하는 것도 지속적으로 이루어져야 함
- AI 시스템은 역동적인 환경에서 지속적으로 진화하고 작동함으로, 신뢰할 수 있는 AI의 실현은 지속적인 프로세스임
- 아래의 방법들은 다른 요구사항과 다른 민감도가 다른 구현 방법에 대한 필요성을 높일 수 있으므로, 서로 보완적이거나 대안이 될 수 있으며, 포괄적이거나 완전한 것도 강제적인 것도 아니고, 그 목적은 신뢰할 수 있는 AI를 구현하는 데 도움이 되는 방법 목록을 제공하는 것임

① 기술적 방법

- 이 부분에서는 AI 시스템의 설계, 개발 및 사용 단계에 통합될 수 있는 신뢰할 수 있는 AI를 보장하는 기술적 방법을 설명하며, 아래 나열된 방법은 성숙도에 따라 다름

<신뢰할 수 있는 AI에 대한 구성>

- 신뢰할 수 있는 AI에 대한 요구사항은 AI 시스템의 구성에 고정되어야 하는 절차 또는 절차에 대한 제약으로 변환되어야 함
- 이는 시스템이 항상 따라야 하는 일련의 '화이트 리스트' 규칙, 시스템이 결코 벗어나서는 안 되는 행위나 상태에 대한 '블랙 리스트' 제한 그리고 그것들의 혼합 또는 그 시스템의 행위에 대한 더 복잡하고 입증 가능한 보증을 통해 성취될 수 있음
- 운영 중 이러한 제한들에 대한 시스템의 준수를 감시하는 것은 별도의 프로세스에 의해 성취될 수도 있음
- 행위를 동적으로 조정할 수 있는 학습 기능을 갖춘 AI 시스템은 예상치 못한 행위을 보이는 비결정적 시스템으로 이해될 수 있음
- 이들은 종종 '감각-계획-행동' 주기의 이론적 렌즈를 통해 고려됨
- 신뢰할 수 있는 AI를 보장하기 위해 이 구성을 적용하는 것은 주기의 세 단계 모두에서 요구사항의 통합이 필요함: ⅰ. '감각' 단계에서 시스템은 요구사항 준수를 보장하는 데 필요한 모든 환경 요소를 인식하도록 개발되어야 함; ⅱ. '계획' 단계에서 시스템은 요구사항을 준수하는 계획만 고려해야 함; ⅲ. '행동' 단계에서 시스템의 작업은 요구사항을 실현하는 동작으로 제한되어야 함
- 위에서 스케치 된 구성은 일반적이며 대부분의 AI 시스템에 대해 불완전한 설명만 제공하지만, 특정 모듈에 반영되어야 하는 제약 및 정책에 관한 기준점을 제공하여 전체 시스템이 신뢰할 수 있고 그렇게 인식되도록 함

<윤리와 디자인에 의한 법(X-by-design)>

- 설계 별 가치를 보장하는 방법은 시스템이 존중해야하는 추상적 원칙과 특정 구현 결정 사이에 정확하고 명시적인 연결을 제공함
- 규범 준수가 AI 시스템 설계에 구현될 수 있다는 생각이 이 방법의 핵심임
- 기업은 처음부터 AI 시스템의 영향과 부정적인 영향을 피하려고 AI 시스템이 준수해야 하는 규범을 식별할 책임이 있음
- 다른 '설계 별' 개념(디자인에 의한 프라이버시와 디자인에 의한 보안)이 이미 널리 사용되고 있음
- 위에서 언급했듯이 신뢰를 얻으려면 AI가 프로세스, 데이터 및 결과에서 안전해야 하며 적대적인 데이터 및 공격에 견고하도록 설계되어야 함
- 안전장치가 되어 있는 종료 메커니즘을 구현하고 강제 종료 후 재개된 작업을

활성화해야 함

<설명 방법>

- 시스템이 신뢰할 수 있으려면 시스템이 특정 방식으로 작동하는 이유와 주어진 해석을 제공한 이유를 이해할 수 있어야 함
- 설명할 수 있는 AI(Explainable AI, XAI)는 시스템의 기본 메커니즘을 더 잘 이해하고 해결책을 찾기 위해 이 문제를 다루려고 하며, 오늘날, 이것은 신경망을 기반으로 하는 AI 시스템에 대한 공공연한 도전임
- 신경망을 사용한 훈련 프로세스는 결과와 연관시키기 어려운 숫자 값으로 설정된 네트워크 매개 변수를 생성할 수 있으며, 때로는 데이터값의 작은 변화로 인해 해석이 크게 변경되어 통학버스를 타조와 혼동하는 것과 같이 시스템을 유도할 수 있음
- 이 취약점은 시스템 공격 중에도 악용될 수 있으며, XAI 연구와 관련된 방법은 사용자에게 시스템의 동작을 설명할 뿐만 아니라 신뢰할 수 있는 기술을 배포하는 데도 중요함

<테스트 및 검증>

- AI 시스템의 비결정적이고 상황에 맞는 특성으로 인해, 기존 테스트로는 충분하지 않음
- 시스템에서 사용하는 개념과 표현의 실패는 프로그램이 충분히 현실적인 데이터에 적용될 때만 나타날 수 있음
- 결과적으로 데이터 처리를 확인하고 검증하기 위해, 이해될 수 있고 예측 가능한 범위 내에서 안정성, 견고성, 작동성을 위한 배치와 훈련 동안 기본 모델이 주의 깊게 모니터되어야만 함
- 계획 프로세스의 결과가 입력과 일치하고 기본 프로세스의 검증이 허락되는 방법으로 결정이 이루어졌는지가 보장되어야만 함
- 시스템의 테스트 및 검증은 시스템이 전체 수명주기 동안 특히 배포 후에 의도한 대로 행동하는지를 보장하기 위해 가능한 한 빨리 이루어져야 함
- 여기에는 데이터, 사전 훈련된 모델, 환경 및 전체 시스템 동작을 포함하여 AI 시스템의 모든 구성요소가 포함되어야 함
- 테스트 프로세스는 가능한 한 다양한 그룹의 사람들이 설계하고 수행해야 하며, 다양한 관점에서 테스트 되는 범주를 다루기 위해 여러 지표를 개발해야 함
- 취약점을 찾기 위해 고의적으로 시스템을 파괴하려는 신뢰할 수 있고 다양한 '레드 팀'[3]의 적대적 테스트와 시스템 오류 및 약점을 감지하고 책임감 있게

3) '레드팀(Red Team)'이란 약점을 공격해 개선 방안을 찾아내는 역할을 부여받은 팀을 말한다.

보고하도록 외부인을 장려하는 “버그 현상금“을 고려할 수 있음
- 마지막으로, 출력 또는 조치가 이전 프로세스의 결과와 일치하는지 확인하고, 이를 이전에 정의된 정책과 비교하여 위반되지 않도록 해야 함

<서비스 지표의 품질>

- 서비스 지표의 적정한 품질은 AI 시스템이 보안과 안전 고려 사항을 염두에 두고 테스트 및 개발되었는지 여부에 대한 기본 이해가 있었는지를 보장하는 것으로서 정의될 수 있음
- 이러한 지표에는 기능, 성능, 유용성, 신뢰성, 보안 및 유지 보수 가능성에 대한 기존 소프트웨어 메트릭뿐만 아니라 알고리즘의 테스트 및 교육을 평가하는 측정이 포함될 수 있음

② 비기술적 방법

- 이 부분에서는 신뢰할 수 있는 AI를 보호하고 유지하는 데 중요한 역할을 할 수 있는 다양한 비기술적 방법을 설명하며, 이것들도 지속해서 평가되어야 함

<규정>
- 제품 안전 법률 및 책임 체계를 볼 때, AI의 신뢰성을 지원하기 위한 규정이 이미 존재함
- 규제를 수정, 조정 또는 도입할 필요가 있다고 생각되면, 보호 수단과 활성화 수단으로, 이는 AI 정책 및 투자 권장 사항으로 구성된 두 번째 결과물에서 제기될 것임

<행동강령>
- 조직과 이해 관계자는 신뢰할 수 있는 AI을 향한 노력을 더하기 위해 지침에 가입하고 기업 책임 헌장, 핵심 성과 지표(“KPI“), 행동강령 또는 내부 정책 문서를 조정할 수 있음
- AI 시스템을 사용하거나 작업하는 조직은 보다 일반적으로 그 의도를 문서화 할 수 있을 뿐만 아니라 기본 권리, 투명성 및 피해 방지와 같은 특정 바람직한 가치의 표준으로 이를 보증할 수 있음

<표준화>

원래는 군사용어로 아군인 블루팀의 약점을 파악하기 위해 편성하는 가상의 적군(敵軍)을 일컫는다. 레드팀은 기존 조직의 시각에서만 판단해 생기는 오류와 피해를 막기 위해 운용된다. <https://www.facebook.com/1483854565240393/posts/1718235485135632/ 참조>.

- 디자인, 제작 및 비즈니스 실행을 위한 표준은 구매 결정을 통해 윤리적 행위를 인식하고 장려하는 능력을 제공함으로써 AI 사용자, 소비자, 조직, 연구 기관 및 정부를 위한 품질 관리 시스템으로 기능 할 수 있음
- 기존 표준 외에도 인증 시스템, 전문 윤리 강령 또는 기본권 준수 디자인 표준과 같은 공동 규제 접근 방식이 존재하며, 현재 예는 ISO 표준 또는 IEEE P7000 표준 시리즈이지만 앞으로는 '신뢰할 수 있는 AI' 라벨이 적합할 수 있으며, 특정 기술 표준을 참조하여 시스템이 안전, 기술적 견고성 및 투명성을 준수하는지 확인함

<인증>

- 모든 사람이 AI 시스템의 작동과 효과를 완전히 이해할 수 있을 것으로 기대할 수 없으므로, AI 시스템이 투명하고 책임 있고 공정하다는 것을 더 많은 대중에게 증명할 수 있는 조직을 고려할 수 있음
- 이러한 인증은 AI 기술을 위해 개발되고, 다양한 환경의 산업 및 사회 표준에 적합하게 조정된 표준을 적용함
- 그러나 인증은 책임을 대체 할 수 없어, 면책 조항, 검토 및 보상 메커니즘을 포함한 책임 체계에 의해 보완되어야 함

<거버넌스 체계를 통한 책무>

- 조직은 내부 및 외부의 거버넌스 체계를 설정하여, AI 시스템의 개발, 배치 및 사용과 관련된 결정의 윤리적 차원에 대한 책무를 보장해야 함
- 여기에는 AI 시스템과 관련된 윤리 문제를 담당하는 사람이나 내부/외부 윤리 패널 또는 이사회의 임명이 포함될 수 있으며, 그러한 사람, 패널 또는 이사회의 가능한 역할 중 하나는 감독과 조언을 제공하는 것임
- 인증 사양 및 기관도 이를 위해 역할을 할 수 있음
- 커뮤니케이션 채널은 업계, 공공 감독 그룹, 모범 사례의 공유, 딜레마의 논의, 윤리 문제들에 대해 나타나는 이슈들을 보고하는 것과 함께 보장되어야만 함
- 이러한 메커니즘은 법적 감독을 보완할 수 있지만 대체 할 수는 없음

<윤리적 사고방식을 촉진하기 위한 교육과 인식>

- 신뢰할 수 있는 AI는 모든 이해 관계자의 정보에 입각한 참여를 장려함
- 커뮤니케이션, 교육 및 훈련은 AI 시스템의 잠재적인 영향에 대한 지식이 널리 퍼져 있는지 확인하고 사람들이 사회 발전을 형성하는 데 참여할 수 있음을 알리는 데 중요한 역할을 함
- 여기에는 모든 이해 관계자가 포함되며, 제품을 만드는 것과 관련된 사람들(디자이너 및 개발자), 사용자(회사 또는 개인) 및 기타 영향을 받는 그룹(AI 시스템을 구매하거나 사용할 수는 없지만, AI 시스템에 의해 결정을 내리는 사람들, 더 넓

게는 사회 전체)이 있음
- 기본 AI 리터러시는 사회 전반에 걸쳐 육성되어야 하며, 대중 교육을 위한 전제 조건은 이 공간에서 윤리학자의 적절한 기술과 훈련을 보장하는 것임

<이해관계인 참여와 사회적 대화>
- AI 시스템의 이점은 많으며, 유럽은 그것들이 모두에게 이용 가능하다는 것을 보장하는 것이 필요함
- 이를 위해서는 공개 토론과 일반 대중을 포함한 사회적 파트너 및 이해 관계자의 참여가 필요하며, 많은 조직이 AI 시스템과 데이터 분석의 사용을 논의하기 위해 이해 관계자 패널에 의존함
- 이 패널에는 법률 전문가, 기술 전문가, 윤리학자, 소비자 대표 및 근로자와 같은 다양한 구성원이 포함되며, AI 시스템의 사용 및 영향에 대해 적극적으로 참여하고 대화를 구하는 것은 결과 및 접근 방식의 평가를 지원하며 특히 복잡한 경우에 도움이 될 수 있음

<다양성과 포용적 디자인 팀>
- 현실 세계에서 사용될 AI 시스템을 개발할 때 다양성과 포용성은 필수적인 역할을 함
- AI 시스템이 자체적으로 더 많은 작업을 수행함에 따라 이러한 시스템을 설계, 개발, 테스트 및 유지 관리, 배치 및 조달하는 팀은 일반적으로 사용자와 사회의 다양성을 반영하는 것이 중요함
- 이것은 객관성과 다양한 관점, 필요 및 목표에 대한 고려에 기여하며, 이상적으로 팀들은 성별, 문화, 연령 뿐만 아니라 직업적 배경 및 기술 측면에서도 다양함

<2장의 주요 가이드>

- 신뢰할 수 있는 AI을 위하여, AI 시스템의 개발, 배치 그리고 이용이 7가지 주요 요건들을 만족하도록 하게 하라
  (1) 인간의 개입 및 감독
  (2) 기술적 견고함과 안전
  (3) 프라이버시와 데이터 거버넌스
  (4) 투명성
  (5) 다양성, 비차별 그리고 공평
  (6) 환경적・사회적 복지
  (7) 책무
- 요건들의 실행을 보장하기 위해 기술적・비기술적 방법들을 고려하라

- AI 시스템 평가를 돕고 그 요건들을 성취하기 위해 연구와 혁신을 촉진하라; 더 많은 공중에게 결과들과 공공연한 문제들을 전파하고, AI 윤리에서 새로운 전문가 세대들을 조직적으로 훈련하라
- 분명하고 사전적인 방법으로, AI 시스템의 능력과 제한, 현실적인 기대 설정 사용 그리고 그 요건들이 실행되는 방법에 대하여 이해관계인들에게 정보를 전달하라
- 그들이 AI 시스템을 다루고 있다는 사실에 대하여 투명해져라
- AI 시스템의 추격 가능성과 감사 가능성을 특히 중요한 환경이나 상황에서 용이하게 하라
- AI 시스템의 전주기 내내 이해관계인들과 관련되라
- 훈련과 교육을 촉진해서 모든 이해관계인들이 신뢰할 수 있는 AI를 알고 훈련되도록 하라
- 서로 다른 원칙과 요건들 사이에서 기본적인 충돌이 존재할 수도 있다는 것에 주의하라
- 이러한 균형과 해결을 계속적으로 정하고, 평가하고, 서면화하고, 전달하라

⑷ 신뢰할 수 있는 AI의 평가

- 2장의 주요 요구사항을 기반으로, 본 장에서는 신뢰할 수 있는 AI를 운영하기 위한 평가 목록(파일럿 버전)을 제시함
- 특히 사용자와 직접 상호 작용하는 AI 시스템에 적용되며, 주로 AI 시스템의 개발자 및 배포자(자체 개발 또는 타사로부터 획득하였던 간에)를 대상으로 함
- 이 평가 목록은 신뢰할 수 있는 AI의 첫 번째 구성요소인 합법적인 AI의 운영을 다루지는 않음
- 이 평가 목록의 준수는 법률 준수의 증거가 아니며 관련 법률 준수를 보장하기 위한 지침으로 의도된 것도 아님
- AI 시스템의 적용 특이성을 고려할 때, 평가 목록은 시스템이 작동하는 특정 사용 사례 및 환경에 맞게 조정되어야 함
- 또한 본 장에서는 운영 및 관리 수준을 모두 포괄하는 거버넌스 구조를 통해 신뢰할 수 있는 AI에 대한 평가 목록을 구현하는 방법에 대한 일반적인 권장 사항을 제공함
- 평가 목록과 거버넌스 구조는 공공 및 민간 부문의 이해 관계자와 긴밀히 협력하여 개발될 것이며, 이 프로세스는 파일럿 프로세스로 구동되어 정성적 프로세스와 정량적 프로세스 두 개의 병렬 프로세스에서 광범위한 피드백을 허용함
- 파일럿 단계가 끝나면 피드백 프로세스의 결과를 평가 목록에 통합하고 2020년 초에 개정된 버전을 준비할 것임
- 목표는 모든 애플리케이션에서 수평적으로 사용할 수 있는 체계를 달성하여 모

든 영역에서 신뢰할 수 있는 AI를 보장하는 기반을 제공하는 것이며, 이러한 기반이 구축되면 분야별 또는 애플리케이션 별 체계를 개발할 수 있음

<거버넌스>

- 이해 관계자는 신뢰할 수 있는 AI 평가 목록을 조직에서 구현할 수 있는 방법을 고려할 수 있음
- 이는 평가 프로세스를 기존 거버넌스 메커니즘에 통합하거나 새로운 프로세스를 구현하여 수행할 수 있음
- 이 선택은 조직의 내부 구조와 규모 및 사용 가능한 리소스에 따라 달라짐
- 연구에 따르면 변화를 달성하기 위해서는 최고 수준의 경영진의 관심이 필수적이라는 사실이 입증되었음
- 또한 회사, 조직 또는 기관의 모든 이해 관계자가 참여하면 새로운 프로세스 도입의 수용 및 관련성을 촉진한다는 사실도 입증되었음
- 따라서 우리는 운영 수준과 최고 경영진의 참여를 모두 포괄하는 프로세스를 구현하는 것을 권장함

<신뢰할 수 있는 AI 평가 리스트를 사용하기>

- 실제로 평가 목록을 사용할 때는 관심 영역뿐 아니라 (쉽게) 답변할 수 없는 질문에도 주의를 기울이는 것이 좋으며, 한 가지 잠재적인 문제는 AI 시스템을 개발하고 테스트하는 팀의 기술과 역량의 다양성이 부족할 수 있으므로 조직 내부 또는 외부의 다른 이해 관계자를 참여시켜야 할 수 있음
- 거버넌스 구조의 모든 수준들에서 문제 해결이 이해될 수 있는 것을 보장하기 위해 모든 결과를 기술적 용어와 관리 용어로 모두 기록하는 것이 강하게 권장됨
- 이 평가 목록은 AI 실무자가 신뢰할 수 있는 AI를 달성하도록 안내하기 위한 것이므로, 평가는 비례하는 방식으로 특정 사용 사례에 맞게 조정되어야 함
- 파일럿 단계에서, 특정 민감한 영역이 드러날 수 있으며 이러한 경우 추가 사양의 필요성은 다음 단계에서 평가됨
- 이 평가 목록은 제기된 질문에 대한 구체적인 답변을 제공하지는 않지만 신뢰할 수 있는 AI가 어떻게 운영될 수 있는지, 그리고 이와 관련하여 취해야 할 잠재적 단계에 대한 반영을 장려함

<기존 법과 프로세스와의 관계>

- AI 실무자들은 특정 프로세스를 의무화하거나 특정 결과를 금지하는 다양한 기존 법률이 있으며, 이는 평가 목록에 나열된 일부 조치와 겹치거나 일치할 수 있음을 인식해야 함
- 예를 들어, 데이터 보호법은 개인 데이터의 수집 및 처리에 관여하는 사람들이

충족해야 하는 일련의 법적 요건을 규정하지만, 신뢰할 수 있는 AI는 데이터의 윤리적 처리도 필요하기 때문에 데이터 보호법 준수를 보장하기 위한 내부 절차 및 정책은 윤리적 데이터 처리를 촉진하는 데 도움이 될 수 있으며, 따라서 기존 법적 프로세스를 보완할 수 있음

- 그러나 이 평가 목록의 준수는 법률 준수의 증거가 아니며 관련 법률 준수를 보장하기 위한 지침이 아님
- 또한 많은 AI 실무자들은 이미 비 법적 표준 준수를 보장하기 위해 기존 평가 도구와 소프트웨어 개발 프로세스를 갖추고 있음
- 아래 평가는 반드시 독립 실행으로 수행되어야 하는 것은 아니지만 이러한 기존 관행에 통합될 수 있음

신뢰할 수 있는 AI 평가 목록 (파일럿 버전)

1. 인간의 개입과 감독

기본권

✓ 기본권에 부정적인 영향을 미칠 수 있는 기본권 영향 평가를 실시 했습니까? 서로 다른 원칙과 권리 사이에서 만들어진 잠재적 균형을 식별하고 문서화 했습니까?

✓ AI 시스템이 인간 (최종) 사용자의 결정과 상호 작용합니까 (예: 권장된 행위 또는 취할 결정, 옵션의 제시)?

- AI 시스템이 의도하지 않은 방식으로 (최종) 사용자의 의사결정 과정을 방해함으로써 인간의 자율성에 영향을 미칠 수 있습니까?
- AI 시스템이 결정, 내용, 조언 또는 결과가 알고리즘 결정의 결과임을 (최종) 사용자에게 전달해야 하는지 여부를 고려 했습니까?
- 채팅 봇 또는 기타 대화 시스템의 경우, 인간의 최종 사용자가 사람이 아닌 기관과 상호 작용하고 있음을 인식합니까?

인간의 개입

✓ AI 시스템이 업무 및 노동 프로세스에 구현되어 있습니까? 그렇다면 의미 있는 상호작용과 적절한 인간 감독 및 제어를 위해 AI 시스템과 인간 간의 작업 할당을 고려했습니까?

- AI 시스템이 인간의 능력을 향상시키거나 증강합니까?
- 업무 프로세스를 위한 AI 시스템에 대한 과신 또는 과잉 의존을 방지하기 위해 보호 조치를 취했습니까?

인간의 감독

✓ 특정 AI 시스템 및 사용 사례에 대한 적절한 수준의 인간 제어를 고려했습니까?

- 인간의 통제 또는 참여 수준을 설명할 수 있습니까?
- “통제중인 인간“은 누구이며 인간 개입의 순간 또는 도구는 무엇입니까?

- 인간의 통제 또는 감독을 보장하기 위한 메커니즘과 조치를 마련했습니까?
- 감사를 활성화하고 AI 자율성 관리와 관련된 문제를 해결하기 위한 조치를 취했습니까?

✓ 자가 학습 또는 자율 AI 시스템 또는 사용 사례가 있습니까? 그렇다면 보다 구체적인 통제 및 감독 메커니즘을 마련했습니까?
- 무언가 잘못될 수 있는지 평가하기 위해 어떤 탐지 및 대응 메커니즘을 설정했습니까?
- 필요한 경우 작업을 안전하게 중단하기 위한 중지 버튼 또는 절차를 확인 했습니까? 이 절차가 프로세스를 전체적으로 또는 부분적으로 중단하거나 제어를 사람에게 위임합니까?

2. 기술적 견고성과 안전성

공격에 대한 탄력 및 보안

✓ AI 시스템이 취약할 수 있는 잠재적 공격 형태를 평가했습니까?
- 데이터 중독, 물리적 인프라, 사이버 공격과 같은 다양한 유형의 취약성을 고려했습니까?

✓ 잠재적 공격에 대한 AI 시스템의 무결성과 복원력을 보장하기 위한 조치 또는 시스템을 마련했습니까?

✓ 예상치 못한 상황과 환경에서 시스템이 어떻게 작동하는지 확인했습니까?

✓ 시스템이 어느 정도 이중 용도로 사용될 수 있는지 고려했습니까? 그렇다면이 사건에 대해 적절한 예방 조치를 취했습니까 (예: 연구를 게시하지 않거나 시스템을 배포하지 않음)?

대체 계획과 일반적인 안전

✓ 적대적 공격이나 기타 예상치 못한 상황(예: 기술 전환 절차 또는 진행하기 전에 작업자에게 요청)이 발생하는 경우 시스템에 충분한 대체 계획이 있는지 확인했습니까?

✓ 특정 사용 사례에서 AI 시스템이 제기하는 위험 수준을 고려했습니까?
- 위험과 안전을 측정하고 평가하기 위한 프로세스를 마련했습니까?
- 신체적 무결성에 대한 위험이 있는 경우 필요한 정보를 제공했습니까?
- AI 시스템의 잠재적인 손상을 처리하기 위해 보험 정책을 고려했습니까?
- 우발적이거나 악의적인 오용을 포함하여 (기타) 예측 가능한 기술 사용의 잠재적인 안전 위험을 식별했습니까? 이러한 위험을 완화하거나 관리 할 계획이 있습니까?

✓ AI 시스템이 사용자나 제3자에게 피해를 줄 가능성이 있는지 평가했습니까? 가능성, 잠재적인 손상, 영향을 받는 청중 및 심각도를 평가했습니까?
- 책임 및 소비자 보호 규칙을 고려하고 이를 고려했습니까?
- 환경이나 동물에 대한 잠재적인 영향이나 안전 위험을 고려했습니까?
- 위험 분석에 사이버 보안 위험과 같은 보안 또는 네트워크 문제가 AI 시스템의 의도하지 않은 동작으로 인해 안전 위험이나 피해를 입힐 수 있는지 여부가 포함되었습니까?

✓ AI 시스템이 잘못된 결과를 제공하거나 사용할 수 없게 되거나 사회적으로 허용할 수 없는 결과(예: 차별)를 제공할 때 AI 시스템의 실패로 인한 영향을 추정했습니까?
- 임계값을 정의하고 대안·대체 계획을 작동시키기 위해 거버넌스 절차를 마련했습니까?
- 대체 계획을 정의하고 테스트했습니까?

정확성

✓ AI 시스템 및 사용 사례의 환경에서 정확도의 수준과 정의가 어느 정도 필요한지 평가했습니까?
- 정확성이 어떻게 측정되고 보장되는지 평가했습니까?
- 사용된 데이터가 포괄적이고 최신인지 확인하기 위한 조치를취했습니까?
- 정확도 향상이나 편견 제거와 같은 추가 데이터가 필요한지를 평가하기 위해 조치를취했습니까?

✓ AI 시스템이 부정확한 예측을 할 경우 어떤 피해가 발생하는지 확인했습니까?

✓ 시스템이 허용할 수 없는 양의 부정확한 예측을 하고 있는지 측정하는 방법을 마련했습니까?

✓ 시스템의 정확성을 높이기 위해 일련의 단계를 마련했습니까?

신뢰성 및 재현성

✓ AI 시스템이 목표, 목적 및 의도 한 적용을 충족하는지 모니터링하고 테스트하는 전략을 세웠습니까?
- 재현성을 보장하기 위해 특정 상황이나 특정 조건을 고려해야 하는지 테스트했습니까?
- 시스템의 신뢰성과 재현성의 다양한 측면을 측정하고 보장하기 위해 검증 방법을 마련했습니까?
- 특정 유형의 설정에서 AI 시스템이 실패할 때를 설명하는 프로세스를 마련했습니까?
- AI 시스템의 신뢰성 테스트 및 검증을 위해 이러한 프로세스를 명확하게 문서화 하고 운영했습니까?
- 시스템의 안정성을 (최종) 사용자에게 보장하기 위해 통신 메커니즘을 설정했습니까?

3. 프라이버시와 데이터 거버넌스

프라이버시 존중 및 데이터 보호

✓ 사용 사례에 따라 AI 시스템의 데이터 수집 (교육 및 운영) 및 데이터 처리 프로세스에서 다른 사람들이 프라이버시 또는 데이터 보호와 관련된 문제를 신고할 수 있는 메커니즘을 설정했습니까?

✓ 데이터 세트의 데이터 유형 및 범위를 평가했습니까 (예: 개인 데이터 포함 여부)?

✓ 잠재적으로 민감한 데이터나 개인 데이터를 사용하지 않거나 최소한으로 사용하여 AI

시스템을 개발하거나 모델을 교육하는 방법을 고려했습니까?

- ✓ 사용 사례에 따라 개인 데이터에 대한 통지 및 제어를 위한 메커니즘을 구축했습니까 (해당되는 경우 유효한 동의 및 취소 가능성 등)?
- ✓ 암호화, 익명화, 집계 등 개인 정보 보호를 강화하기 위한 조치를 취했습니까?
- ✓ 데이터 개인 정보 보호 책임자(DPO)가 있는 경우, 프로세스 초기 단계에서 이 사람을 참여시켰습니까?

데이터의 품질과 무결성

- ✓ 시스템을 관련 표준(예: ISO, IEEE) 또는 일상적인 데이터 관리 및 거버넌스를 위해 널리 채택된 프로토콜에 맞게 조정 했습니까?
- ✓ 데이터 수집, 저장, 처리 및 사용에 대한 감독 메커니즘을 설정했습니까?
- ✓ 사용된 외부 데이터 소스의 품질을 어느 정도 관리하고 있는지 평가했습니까?
- ✓ 데이터의 품질과 무결성을 보장하기 위한 프로세스를 마련했습니까? 다른 프로세스를 고려했습니까? 데이터 세트가 손상되거나 해킹되지 않았는지 어떻게 확인하고 있습니까?

데이터에 대한 접근

- ✓ 적절한 데이터 거버넌스를 관리하고 보장하기 위해 어떤 프로토콜, 프로세스 및 절차를 따랐습니까?
  - 누가 사용자의 데이터에 액세스할 수 있으며 어떤 상황에서 액세스할 수 있는지 평가했습니까?
  - 이러한 사람이 데이터에 액세스할 수 있는 자격을 갖추고 있으며 데이터 보호 정책의 세부 사항을 이해하는 데 필요한 능력이 있는지 확인했습니까?
  - 언제, 어디서, 어떻게, 누가, 어떤 목적으로 데이터에 액세스했는지 기록하는 감독 메커니즘을 확보했습니까?

4. 투명성

추적 가능성

- ✓ 추적성을 보장할 수 있는 조치를 설정했습니까? 이를 위해 다음 방법을 문서화 해야 합니다.
  - ➢ 알고리즘 시스템 설계 및 개발에 사용되는 방법:
    - o 규칙 기반 AI 시스템 : 프로그래밍 방법 또는 모델 구축 방법
    - o 학습 기반 AI 시스템; 어떤 입력 데이터가 수집되고 선택되었는지, 이것이 어떻게 발생했는지를 포함한 알고리즘 훈련 방법
  - ➢ 알고리즘 시스템을 테스트하고 검증하는 데 사용되는 방법:
    - o 규칙 기반 AI 시스템 테스트 및 검증에 사용되는 시나리오 또는 사례
    - o 학습 기반 모델 : 테스트 및 검증에 사용되는 데이터에 대한 정보.
  - ➢ 알고리즘 시스템의 결과:
    - o 알고리즘에 의해 취해진 결과 또는 결정과 다른 경우 (예: 다른 사용자 하위

그룹의 경우)에서 발생할 수 있는 잠재적인 기타 결정.

설명 가능성

✓ 다음을 평가했습니까?
- AI 시스템이 내린 결정과 결과를 어느 정도까지 이해할 수 있습니까?
- 시스템의 결정이 조직의 의사결정 프로세스에 어느 정도 영향을 미칩니까?
- 이 특정 시스템이 이 특정 영역에 배치된 이유는 무엇입니까?
- 시스템의 비즈니스 모델은 무엇입니까(예: 조직의 가치를 창출하는 방법)?

✓ 시스템이 특정 선택을 하여 모든 사용자가 이해할 수 있는 특정 결과를 얻은 이유를 설명했습니까?

✓ 처음부터 해석 가능성을 염두에 두고 AI 시스템을 설계했습니까?
- 해당 응용 프로그램에 대해 가능한 가장 간단하고 해석 가능한 모델을 조사하고 사용하려고 했습니까?
- 훈련 및 테스트 데이터를 분석할 수 있는지 평가했습니까? 시간이 지남에 따라 이를 변경하고 업데이트할 수 있습니까?
- 모델의 교육 및 개발 후 해석 가능성을 검사 할 수 있는지 또는 모델의 내부 워크플로에 액세스할 수 있는지 평가했습니까?

커뮤니케이션

✓ (최종) 사용자에게-면책 조항이나 다른 수단을 통해-그들이 다른 사람이 아니라 AI 시스템과 상호 작용하고 있다는 것을 전달했습니까? AI 시스템에 그렇게 라벨을 붙였습니까?

✓ (최종) 사용자에게 AI 시스템 결과의 이유와 기준을 알리는 메커니즘을 설정 했습니까?
- 의도된 청중에게 이것을 명확하고 이해하기 쉽게 전달했습니까?
- 사용자의 피드백을 고려하고 이를 사용하여 시스템을 조정하는 프로세스를 설정했습니까?
- 편견과 같이 잠재적이거나 인지된 위험에 대해 의사소통했습니까?
- 사용 사례에 따라 다른 청중, 제3자 또는 일반 대중에 대한 커뮤니케이션 및 투명성을 고려했습니까?

✓ AI 시스템의 목적과 제품・서비스의 혜택을 누릴 수 있는 사람 또는 무엇을 명확히 하셨나요?
- 제품에 대한 사용 시나리오를 지정하고 의도 한 대상에 대해 이해 가능하고 적절하도록 명확하게 전달했습니까?
- 사용 사례에 따라 인간 심리학과 혼란의 위험, 확증 편향 또는 인지 피로와 같은 잠재적인 한계에 대해 생각했습니까?

✓ AI 시스템의 특성, 한계 및 잠재적인 단점을 명확하게 전달했습니까?
- 시스템 개발의 경우: 제품이나 서비스에 배포하는 사람에게?
- 시스템 배포의 경우: (최종) 사용자 또는 소비자에게?

5. 다양성, 비차별 그리고 공정성

불공정한 편견의 회피

✓ 입력 데이터 사용과 알고리즘 설계 모두와 관련하여 AI 시스템에서 불공정한 편견을 만들거나 강화하지 않도록 전략 또는 절차를 설정했습니까?
- 사용된 데이터 세트의 구성으로 인한 가능한 제한을 평가하고 인정했습니까?
- 데이터에서 사용자의 다양성과 대표성을 고려했습니까? 특정 모집단이나 문제가 있는 사용 사례를 테스트했습니까?
- 데이터, 모델 및 성능에 대한 이해를 높이기 위해 사용 가능한 기술 도구를 조사하고 사용했습니까?
- 시스템의 개발, 배포 및 사용 단계에서 잠재적인 편견을 테스트하고 모니터링하는 프로세스를 마련했습니까?

✓ 사용 사례에 따라 다른 사람들이 AI 시스템의 편견, 차별 또는 성능 저하와 관련된 문제를 신고할 수 있는 메커니즘을 확보했습니까?
- 그러한 문제를 누구에게 어떻게 제기할 수 있는지에 대한 명확한 단계와 의사소통 방법을 설정했습니까?
- (최종) 사용자 외에 AI 시스템의 잠재적인 간접적인 영향을 받는 다른 사람을 고려했습니까?

✓ 동일한 조건에서 발생할 수 있는 의사결정 변동성이 있는지 평가했습니까?
- 그렇다면, 가능한 원인이 무엇인지 고려했습니까?
- 변동성의 경우, 그러한 변동성이 기본권에 미치는 잠재적 영향에 대한 측정 또는 평가 메커니즘을 설정했습니까?

✓ AI 시스템을 설계할 때 적용하는 “공정성“에 대한 적절한 작업 정의를 보장했습니까?
- 귀하의 정의가 일반적으로 사용됩니까? 이것을 선택하기 전에 다른 정의를 고려했습니까?
- 적용된 공정성 정의를 측정하고 테스트하기 위해 정량적 분석 또는 메트릭을 확인했습니까?
- AI 시스템의 공정성을 보장하기 위한 메커니즘을 설정했습니까? 다른 잠재적 메커니즘을 고려했습니까?

접근성 및 보편적인 디자인

✓ AI 시스템이 다양한 개인 취향과 능력을 수용하는지 확인했습니까?
- 특별한 도움이 필요하거나 장애가 있는 사람이나 배제 위험이 있는 사람이 AI 시스템을 사용할 수 있는지 평가했습니까? 이것이 시스템에 어떻게 설계되었으며 어떻게 확인됩니까?
- AI 시스템에 대한 정보가 보조기술 사용자도 액세스할 수 있는지 확인했습니까?
- AI 시스템 개발 단계에서 이 커뮤니티에 참여했거나 상담했습니까?

✓ AI 시스템이 잠재적인 사용자 청중에게 미치는 영향을 고려했습니까?
- AI 시스템 구축에 참여한 팀이 대상 사용자를 대표하는지 평가했습니까? 살짝 영향을 받을 수 있는 다른 그룹을 고려할 때 더 넓은 인구를 대표합니까?

- 부정적인 의미로 인해 불균형적으로 영향을 받을 수 있는 개인이나 집단이 있는지 평가했습니까?
- 배경과 경험이 다른 팀이나 그룹으로부터 피드백을 받았습니까?

이해 관계자 참여

✓ AI 시스템의 개발 및 사용에 다양한 이해 관계자의 참여를 포함하는 메커니즘을 고려했습니까?

✓ 영향을 받는 근로자와 그 대표자들에게 사전에 알리고 참여시켜 조직에 AI 시스템을 도입할 수 있는 길을 닦았습니까?

6. 사회적 그리고 환경적 복지

지속 가능하고 환경 친화적인 AI

✓ AI 시스템의 개발, 배포 및 사용이 환경에 미치는 영향을 측정하는 메커니즘을 설정했습니까(예: 데이터 센터에서 사용하는 에너지 유형)?

✓ AI 시스템의 수명주기가 환경에 미치는 영향을 줄이기 위한 조치를 취했습니까?

사회적 영향

✓ AI 시스템이 인간과 직접 상호 작용하는 경우:

- AI 시스템이 인간이 시스템에 대한 애착과 공감을 개발하도록 장려하는지 평가했습니까?
- 사회적 상호 작용이 시뮬레이션 되고 “이해“와 “느낌“의 능력이 없다는 것을 AI 시스템이 분명하게 나타냈는지를 확인했습니까?

✓ AI 시스템의 사회적 영향을 잘 이해하고 있는지 확인했습니까? 예를 들어, 인력의 실직 또는 기술 감퇴 위험이 있는지 평가했습니까? 그러한 위험에 대응하기 위해 어떤 조치를 취했습니까?

사회와 민주주의

✓ 간접적으로 영향을 받을 가능성이 있는 이해 관계자와 같이 개인 (최종) 사용자를 넘어서 AI 시스템 사용이 사회에 미치는 광범위한 영향을 평가했습니까?

7. 책무

감사 가능성

✓ AI 시스템의 프로세스 및 결과에 대한 추적 및 기록과 같이 시스템의 감사 가능성을 촉진하는 메커니즘을 설정했습니까?

✓ 기본 권한에 영향을 미치는 애플리케이션(안전에 중요한 애플리케이션 포함)에서 AI 시스템이 독립적으로 감사 될 수 있는지 확인했습니까?

부정적인 영향의 최소화 및 보고

✓ (간접) 영향을 받는 다양한 이해 관계자를 고려하여 AI 시스템에 대한 위험 또는 영향 평가를 수행했습니까?

✓ 책임 관행 개발을 돕기 위해 훈련과 교육을 제공했습니까?

- 팀의 어떤 직원이나 지점이 관련되어 있습니까? 개발 단계를 넘어가나요?
- 이러한 교육은 AI 시스템에 적용할 수 있는 잠재적인 법적 체계도 가르치는가?
- 잠재적으로 불명확한 회색 영역을 포함하여 전반적인 책임 및 윤리 관행을 논의하기 위해 '윤리적 AI 검토위원회' 또는 유사한 메커니즘을 구축하는 것을 고려했습니까?

✓ 내부 이니셔티브 외에도 윤리 및 책임을 감독하기 위한 외부 지침을 예상했거나 감사 프로세스를 마련했습니까?

✓ 제3자(예: 공급 업체, 소비자, 유통 업체・벤더) 또는 근로자가 AI 시스템의 잠재적인 취약성, 위험 또는 편견을 보고 할 수 있는 프로세스를 설정했습니까?

균형의 기록

✓ AI 시스템과 관련된 관심사와 가치, 그리고 그들 사이의 잠재적인 절충점을 식별하는 메커니즘을 설정했습니까?

✓ 그러한 균형을 어떻게 결정합니까? 절충 결정이 문서화 되었는지 확인했습니까?

보상의 능력

✓ 피해나 악영향이 발생할 경우 시정 할 수 있는 적절한 메커니즘을 설정했습니까?

✓ 교정 기회에 대한 정보를 (최종) 사용자나 제3자에게 제공하는 메커니즘을 모두 마련했습니까?

자가 평가를 위한 신뢰할 수 있는 AI를 위한 평가 리스트(개선)

- The Assessment List for Trustworthy AI for Self Assessment(ALTAI) -

1. 인간의 개입과 감독
 - 인간의 개입과 자율성
 - 인간의 감독
2. 기술적 강건함과 안전
 - 공격에 대한 탄력 및 보안
 - 일반 안전
 - 정확성
 - 신뢰성, 대체 계획과 재현성
3. 프라이버시와 데이터 거버넌스
 - 프라이버시
 - 데이터 거버넌스
4. 투명성
 - 추적가능성

- 설명가능성
- 커뮤니케이션

5. 다양성, 비차별, 형평성
- 불공정한 편견의 회피
- 접근성과 보편적 디자인
- 이해관계인의 참가

6. 사회적・환경적 복지
- 환경적 복지
- 직장과 기술에 대한 영향
- 더 넓은 사회나 민주주의에 대한 영향

7. 책무
- 감사 가능성
- 위험 관리

<3장의 주요 가이드>

- AI 시스템을 개발, 배치 또는 이용할 때 신뢰할 수 있는 AI 평가 리스트를 채택하고, 그 시스템이 적용되고 있는 구체적인 사용 사례에 그것을 조정하라
- 그러한 평가 리스트는 결코 완전할 것이지 않다는 것을 명심하라
- 신뢰할 수 있는 AI를 만드는 것은 박스에 체크하는 것에 대한 것이 아니고 요건들을 계속해서 정의하고 실행하고, 해결책을 평가하고 개선된 결과들을 보증하고, 이러한 식으로 이해관계인들과 관련되도록 하는 것에 대한 것이다

4. 결론

- 신뢰할 수 있는 AI는 3가지 구성요소를 가짐: (1) 합법적이어야 함; (2) 윤리적이어야 함; (3) 견고해야 함
- 1장에서는 AI 환경에서 중요한 기본권과 상응하는 윤리적 원칙들을 설명함
- 2장에서는 신뢰할 수 있는 AI를 실현하기 위해 AI 시스템들이 만족해야만 하는 7개의 주요 요건들을 열거하고, 그들의 이행을 도울 수 있는 기술적・비기술적 방법들을 제안하였음
- 3장에서는 7가지 요건들을 운용하는데 도움을 줄 수 있는 신뢰할 수 있는 AI 평가 리스트를 제공하였음
- 마지막 부분에서는 유익한 기회의 예와 AI 시스템에 의해 제기된 중요한 문제 들을 제공하였음

Ⅲ. 시사점

- 국내 AI 윤리 헌장 중 카카오의 ‘알고리즘 윤리 헌장’, 한국인공지능윤리협회(KAIEA)의 ‘인공지능 윤리 헌장’, 국토교통부의 ‘자율주행차 가이드라인’, 한국정보화진흥원(NIA)의 ‘지능정보사회 윤리헌장’ 등이 있음
- 국회 입법조사처는 ‘AI의 윤리적 사용을 위한 개선과제’ 보고서에서 AI의 안전한 사용을 위해 AI 설계과정과 최초 개발 목적대로 사용하고 있는지를 점검할 수 있는 기준의 마련이 필요하다고 함
- 입법조사처[4]는 다음과 같은 개선책을 제시함: 윤리기준과 관련된 문제를 조정·해결할 수 있는 거버넌스를 둬야함; AI 사용에 따른 문제점을 사후에 추적·평가할 수 있는 제도마련이 필요함; 대학과 대학원, 일반 시민을 대상으로 한 AI 윤리교육 도입, 기업의 AI 윤리 책임 강화, AI 피해에 대한 배상책임 제도 보완 등이 필요함
- AI 국가전략은 안전한 AI 사용을 위한 AI 역기능 방지와 AI 윤리 정립을 제안하였지만, EU 가이드라인과 같이 AI의 주된 목적(AI 윤리의 제목), 구성요소, 근거, 실현을 위한 요건과 방법, 그리고 구체적인 평가방법까지를 제공하고 있지는 못함
- 향후 AI 관련 산업의 윤리적 기준과 원칙으로서 작용할 종합적이고 구체적이며 그리고 우리에게 적합한 표준화된 가이드라인이 필요함

- EU 가이드라인의 가정적 적용

자율주행차 윤리가이드라인

• 목표와 기본 가치 *→ 제목으로서 ‘신뢰할 수 있는 인공지능’ 또는 ‘인간 중심의 인공지능’(목표와 제목의 일치)*
  - 자율주행차의 목표는 인간의 안전과 복리 증진
  - 인간의 안전하고 편리하며 자유로운 이동권 보장
  - 인간의 생명을 동물이나 재산의 피해보다 우선적으로 고려(인간의 존엄성)
  - 사고로 인한 개인적, 사회적 손실의 최소화(공공선)

  → 근거: 인간의 존엄성에 대한 존경, 개인의 자유, 민주주의와 공정성 그리고 법치에 대한 존중, 평등과 비차별 및 연대, 시민의 권리

• 행위 준칙
  - 투명성, 제어가능성, 책임성, 안전성, 보안성 *→ 윤리적 원칙: 인간의 자율성에 대한 존중, 피해 방지, 공정성, 설명 가능성 원칙*

  → AI의 요건: 인간의 개입과 감독, 기술적 견고성과 안전성, 개인정보보호 및 데이터 거버넌스, 투명성, 다양성과 차별금지 및 공정성, 사회적·환경적 복지, 책무(각 요건에서의 구체적 요구 조건들)

• 자율주행차와 관련된 각 주체들이 준수해야하는 윤리 원칙

4) 이순기, 인공지능의 윤리적 사용을 위한 개선과제, 이슈와 논점 제1759호.

- 설계자의 의무: 자율차를 불법 개조하거나 임의로 시스템을 변경할 수 없도록 시스템 설계, 해킹을 방지할 수 있도록 자율차 설계 등
- 제작자의 의무: 제작・판매에 관련된 법규 준수, 자율차의 안전과 보안에 대한 보장 책임, 사용연한 내의 유지보수와 결함에 대한 책임 등
- 관리자의 의무: 자율차 도입과 활용을 위한 사회적 인프라 확충 의무, 자율차 도입, 안전 및 모니터링 등에 관한 의무 등
- 소비자의 의무: 자율차 임의 개조・변경 금지, 오사용 및 불법적 사용으로 발생하는 문제에 대한 책임 의무, 법률 및 사용지침 준수 등

→ 실현을 위한 기술적・비기술적 방법: 구성, 디자인, 설명방법, 테스트 및 검증, 서비스 지표의 품질 / 규정, 행동강령, 표준화, 인증, 거버넌스 체계, 교육과 인식, 이해관계인의 참여, 다양성과 표용적 디자인

→ 평가 목록: 기본권 영향 평가를 실시했습니까? 인간의 감독 및 제어가 가능합니까? 인간 사용자의 결정과 상호 작용합니까? 감재적 공격 형태를 평가했습니까? 무결성과 복원력을 보장하기 위한 조치를 마련했습니까?

→ 목표, 원칙, 실현방법, 평가방법은 추후 AI의 설계, 제작, 관리, 소비의 기준이나 방법이 될 것이며, 평가나 인증 등의 기준으로 자리잡을 수 있음