FPRI
Future Policy Research Institute

# "AI & Ethics Guidelines"

- 인공지능과 윤리가이드라인 V1.0 -

2019.07.21

Jason, Min

## □ 1942 아이작아시모프 – 로봇 3+1 원칙

ㅇ.1원칙: A robot may not injure a human being or, through inaction, allow a human being to come to harm(로봇은 인간에게 해를 끼치거나, 아무런 행동도 하지 않음으로써 인간에게 해가 가도록 해서는 안 된다).

ㅇ.2원칙: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law(로봇은 인간의 명령에 복종해야 한다. 단 명령이 첫 번째 원칙에 위배될 때는 예외로 한다).

ㅇ.3원칙: A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws(로봇은 자신을 보호해야 한다. 단 첫 번째와 두 번째 원칙과 위배될 때는 예외로 한다)

1985년, 아시모프의 로봇 0원칙

'A robot may not harm humanity, or, by inaction, allow humanity to come to harm'(로봇은 인류에게 해를 가하거나, 행동을 하지 않음으로써 인류에게 해가 가도록 해서는 안 된다).

https://brunch.co.kr/@kakao-it/184

## □ 2004년, 일본 후쿠오카 세계 로봇 선언

- 차세대 로봇은 인간과 공존하는 파트너가 될 것이다.
- 차세대 로봇은 인간을 육체적으로 그리고 정신적으로 보조할 것이다.
- 차세대 로봇은 안전하고 평화로운 사회 구현에 기여할 것이다

https://brunch.co.kr/@kakao-it/184

## □ 2006년, 유럽로봇연구연합(EURON)의 로봇윤리 로드맵

[로봇윤리 원칙]

- 인간의 존엄과 인간의 권리
- 평등, 정의, 형평
- 편익과 손해
- 문화적 다양성을 위한 존중
- 차별과 낙인화의 금지
- 자율성과 개인의 책무성
- 주지된 동의
- 프라이버시
- 기밀성
- 연대와 협동
- 사회적 책무
- 이익의 공유
- 지구 상의 생물에 대한 책무

https://brunch.co.kr/@kakao-it/184

## □ 2007년, 대한민국 산업자원부의 로봇윤리헌장

[산업자원부 로봇윤리헌장(초안)]

1장(목표) 로봇윤리헌장의 목표는 인간과 로봇의 공존공영을 위해 인간중심의 윤리규범을 확인하는 데 있다.

2장(인간, 로봇의 공동원칙) 인간과 로봇은 상호간 생명의 존엄성과 정보, 공학적 윤리를 지켜야 한다.

3장(인간 윤리) 인간은 로봇을 제조하고 사용할 때 항상 선한 방법으로 판단하고 결정해야 한다.

4장(로봇 윤리) 로봇은 인간의 명령에 순종하는 친구・도우미・동반자로서 인간을 다치게 해서는 안 된다.

5장(제조자 윤리) 로봇 제조자는 인간의 존엄성을 지키는 로봇을 제조하고 로봇 재활용, 정보보호 의무를 진다.

6장(사용자 윤리) 로봇 사용자는 로봇을 인간의 친구로 존중해야 하며 불법개조나 로봇남용을 금한다.

7장(실행의 약속) 정부와 지자체는 헌장의 정신을 구현하기 위해 유효한 조치를 시행해야 한다.

https://brunch.co.kr/@kakao-it/184

## □ 2010년, 영국의 공학과 물리과학 연구위원회의 로봇 원칙

[EPSRC의 로봇 윤리]

ㅇ.로봇은 단 국가안보를 위한 경우는 제외하고, 무기로서 설계되어서는 안된다.

ㅇ.로봇은 사생활 보호 뿐 아니라 프라이버시를 포함한 현존하는 법규범에 부합되도록 설계되고 운용되어야 한다.

ㅇ.로봇은 안전과 보안 보장에 적합하도록 설계되어야 한다.

ㅇ.로봇은 사람에게 이에 대한 착각이나 환상을 불러 일으키도록 설계되고 사용되어서는 안된다.

ㅇ.모든 로봇에 대해 법적 책임을 지는 사람이 명확하게 명시되어야 한다.

https://brunch.co.kr/@kakao-it/184

## □ 2016년 4월, 일본의 AI R&D 가이드라인

[ 일본의 AI R&D 가이드라인 ]

ㅇ투명성: AI 네트워크 시스템을 설명하고 검증할 수 있는 능력 확보

ㅇ이용자 지원: AI 네트워크 시스템이 이용자를 지원하고 이용자에게 선택의 기회를 적절히 제공

ㅇ통제: 인간에 의한 AI 네트워크 시스템의 통제 가능성 보장

ㅇ보안: AI 네트워크 시스템의 견고성 및 신뢰성 확보

ㅇ안전: AI 네트워크 시스템이 사용자 및 제 3 자의 생명 또는 신체에 위험을 초래하지 않도록 고려

ㅇ프라이버시: AI 네트워크 시스템이 사용자 및 제 3 자의 개인 정보를 침해하지 않도록 고려

ㅇ윤리: 네트워크화 될 인공 지능의 연구 개발을 수행함에 있어 인간의 존엄성과 개인의 자율성 존중

ㅇ책임: AI 연구원 또는 개발자와 네트워크를 통해 사용자와 같은 이해 관계자에 대한 책임성 부여

https://brunch.co.kr/@kakao-it/184

## □ 2016년 6월, 사티아 나델라의 인공지능 규칙

[사티아 나델라의 인공지능 규칙]

ㅇ 인공지능은 인간을 돕기 위해 개발되어야 한다.

ㅇ 인공지능은 투명해야 한다.

ㅇ 인공지능은 인간의 존엄성을 파괴하지 않으면서 효율을 극대화해야한다.

ㅇ 인공지능은 개인정보 보호를 위해 설계 되어야한다.

ㅇ 인공지능은 의도하지 않은 피해를 인간이 복구할 수 있도록 알고리즘 차원의 설명 책임을 지닌다.

ㅇ 인공지능은 차별 또는 편견을 방지해야 한다.

https://brunch.co.kr/@kakao-it/184

## □ 2016년, 한국의 로봇윤리헌장 개선안

1.1. 로봇은 인간의 존엄성을 존중해야 한다.
1.2. 로봇은 인류의 공공선을 실현하는 데 기여해야 한다.
1.3. 로봇은 인류의 공공선을 침해하지 않는 범위 내에서 인간의 존엄성을 추구해야 한다.
1.4. 로봇은 위의 세 원칙들을 위배하지 않는 범위 내에서 사용자의 명령을 준수해야 한다.
1.5. 로봇은 위 원칙들을 준수해야 하며, 이에 대한 책임은 설계 및 제작자에게 있다.
1.6. 로봇은 설계 및 제작의 목적에 부합하여 사용되어야 하며, 그 이외의 사용에 대한 책임은 사용자에게 있다.
2.0. 로봇설계자는 인간의 존엄성 준중과 인류의 공공선 실현에 기여하는 로봇을 설계해야 한다.
2.1. 로봇 설계자는 헌법에 명시된 인간의 기본권을 침해하지 않도록 설계해야 한다.
2.2. 로봇 설계자는 개인을 포함한 공동체 전체의 선을 침해하지 않도록 설계해야 한다.
2.3. 로봇 설계자는 생태계를 포함한 생명 공동체의 지속가능성을 침해하지 않도록 설계해야 한다.
2.4. 로봇 설계자는 정보통신 윤리 및 기술, 공학 윤리와 관련된 강령들을 준수해야 한다.
2.5. 로봇 설계자는 로봇의 목적 및 기능을 설정하고 이에 맞게 사용되도록 설계해야 한다.
3.0. 로봇 제작자는 공익의 범위 내에서 인간의 행복추구에 도움이 되고 정해진 목적과 기능에 부합하는 로봇을 만들어야 한다.
3.1. 로봇 제작자는 로봇의 제작 및 판매와 관련된 법규를 준수하여 제작해야 한다.
3.2. 로봇 제작자는 로봇의 목적 및 기능과 관련된 법규나 인증에 따라 로봇을 제작해야 한다.
3.3. 로봇 제작자는 설계된 로봇의 목적 및 기능을 변경하지 않고 제작해야 한다.
3.4. 로봇 제작자는 로봇의 사용 연한을 정하고 폐기에 대한 지침을 제공해야 한다.
3.5. 로봇 제작자는 로봇의 사용 연한 내에서의 유지보수와 결함으로 발생된 피해에 대한 책임을 져야 한다.
4.0. 로봇 사용자는 자신이나 타인의 삶의 질과 복지의 향상을 위해 정해진 목적과 기능에 따라 로봇을 사용해야 한다.
4.1. 로봇 사용자는 로봇 사용과 관련된 법률과 사용 지침을 준수해야 한다.
4.2. 로봇 사용자는 로봇을 불법적으로 개조하거나 임의로 변경할 수 없으며, 정해진 목적 및 기능에 맞게 사용해야 한다.
4.3. 로봇 사용자는 로봇을 사용하여 타인의 이익을 침해하거나 위해를 가하지 않도록 해야 한다.
4.4. 로봇 사용자는 로봇의 오용 및 불법적 사용으로 발생하는 문제에 대하여 책임을 져야 한다.
4.5. 로봇 사용자는 로봇 남용으로 발생할 수 있는 과몰입, 의존, 중독 등에 주의해야 한다.

https://brunch.co.kr/@kakao-it/184

## □ 2017년, 경제협력개발기구(OECD) - 규범적 대응원칙 및 정책방향

AI 자동화에 따른 미래 일자리 변화에 대한 정책적 대응

프로그램 개발과 비판적 사고 및 공감 능력 등 다양한 AI 활용 역량 증진의 중요성

알고리즘 차별 등 인간에 영향 미칠 수 있는 알고리즘 의사 결정 문제에 대해 투명성

공정성

책무성의 가치 개입

https://brunch.co.kr/@kakao-it/184

FPRI
Future Policy Research Institute

## □ 2017.1.17 Asilomar AI Principle (아실로마 인공지능 원칙)
## * 비영리 연구 단체 Future of Life Institute

미국 캘리포니아의 아실로마에서 열린 AI 콘퍼런스(06~08 Jan 2017)에서 23개 원칙 도출

* 테슬라의 일론 머스크, 알파고 창시자 데미스 허사비스, 과학자 스티븐 호킹 등이 서명

연구이슈(Research Issues) 5개
윤리와 가치(Ethics and Value) 13개
장기적 이슈(Longer-term Issues) 5개

https://futureoflife.org/ai-principles/?cn-reloaded=1

## 연구 이슈 (Research Issues)

1) **연구 목표** : 인공 지능(AI) 연구의 목표는 지향하는 바가 없는 지능이 아니라 **유익한 지능을 창출**하는 것입니다.

2) **연구비 지원** : AI에 대한 투자에는 다음과 같이 컴퓨터 과학, 경제, 법, 윤리 및 사회 연구 등의 어려운 질문을 포함한, 유익한 사용을 보장하는 연구를 위한 기금이 동반되어야 합니다.
•미래의 인공 지능 시스템이 **오작동이나 해킹** 없이 우리가 원하는 것을 수행할 수 있도록 매우 탄탄하게 만들 수 있는 방안은 무엇입니까?
•인류의 자원과 목적을 유지하면서 자동화를 통해 우리가 계속 번영할 수 있는 방안은 무엇입니까?
•AI와 보조를 맞추고  그와 관련된 위험을 관리하기 위해 법률 시스템을  보다 공정하고 효율적으로 업데이트 할 수 있는 방안은 무엇입니까?
•AI는 어떠한 가치들에 따라야 하며, 그것이 가져야 하는 법적, 윤리적 상태는 무엇입니까?

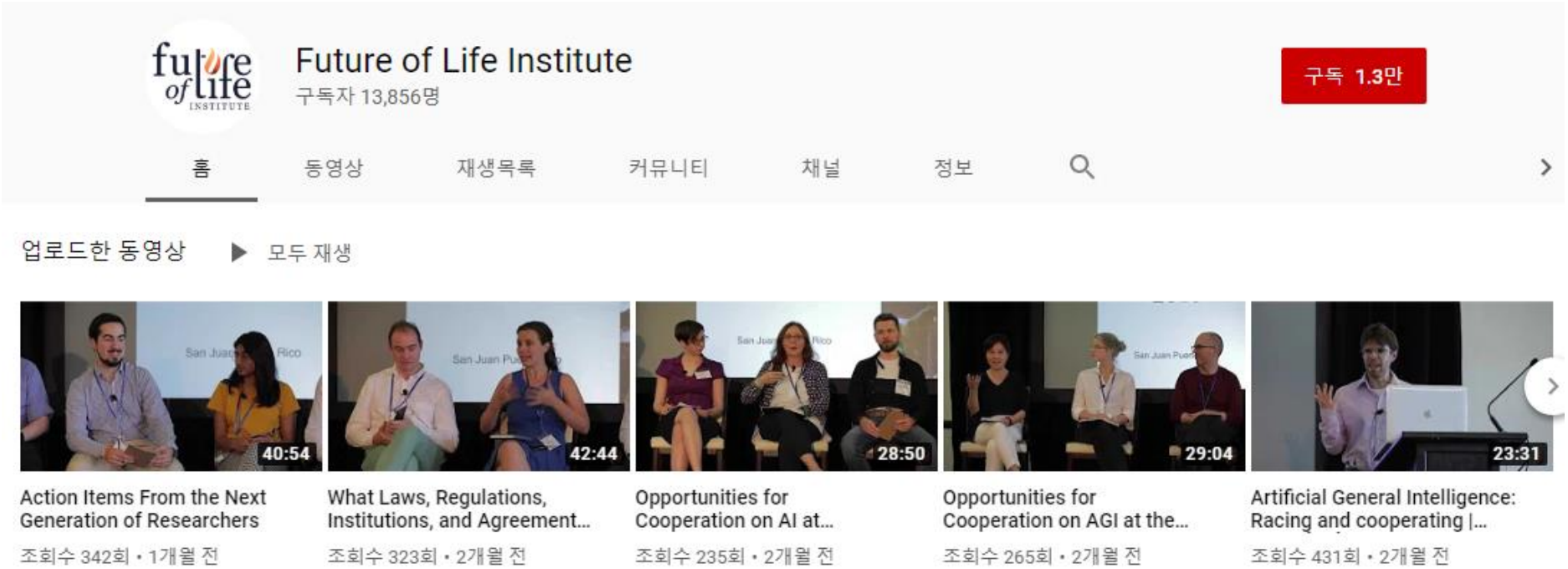
3) **과학-정책 관계** : AI 연구자와 정책 입안자간에 건설적이고 건전한 교류가 있어야 합니다.

4) **연구 문화** :AI 의 연구자와 개발자간에 협력, 신뢰, 투명성의 문화가 조성되어야 합니다.

5) **경쟁 회피**: AI 시스템을 개발하는 팀들은 안전 기준에 대한 질 낮은 해결책을 피하기 위해 적극적으로 협력해야 합니다.

FPRI
Future Policy Research Institute

https://www.youtube.com/channel/UC-rCCy3FQ-GItDimSR9lhzw



## 윤리와 가치 (Ethics and Values)

6) **안전** : AI 시스템은 작동 수명 전반에 걸쳐 안전하고 안정적이어야 하며, 적용과 실현이 가능하다면  검증 할 수 있어야 합니다.

7) **오류 투명성**: AI 시스템이 해를 입히는 경우 그 이유를 확인할 수 있어야 합니다.

8) **사법의 투명성** : 사법 결정에 있어 자동화된 시스템이 개입할 경우,  권한있는 인간 기관이 감사할 수 있는  충분한 설명을 제공해야 합니다.

9) **책임성**: 고급 AI 시스템의 설계자와 제조자는  그것의 사용, 오용 및 행위의 도덕적 함의에 있어서,  그것을 형성할 책임과 기회가 있는 이해관계자 입니다.

10) **가치의 준수** : 고도로 자율적인 AI 시스템은 그것이 작동하는 동안 목표와 행동이 인간의 가치와 반드시 일치하도록 설계되어야 합니다.

11) **인간의 가치** : AI 시스템은 인간의 존엄성, 권리, 자유 및 문화 다양성의 이상과 양립할 수 있도록 설계되고 운영되어야 합니다.

12) **개인 정보 보호** : AI 시스템이 개인정보 데이터를 분석하고 활용할 수 있는 경우, 사람들은 자신이 생성한 데이터에 접근해 관리 및 제어할 권리를 가져야 합니다.

13) **자유와 개인 정보** : 개인정보에 대한 AI 의 적용이 사람들의 실제 또는 인지된 자유를 부당하게 침해해서는 안됩니다.

14) **이익 공유** : AI  기술은 가능한 많은 사람들에게 혜택을 주고 역량을 강화해야 합니다.

15) **공동 번영** : AI에 의해 만들어진 경제적 번영은 모든 인류에게 이익이 되도록 널리 공유되어야 합니다.

16) **인간 통제** : 인간은 인간이 선택한 목적를 달성하기 위해, 의사 결정을 AI 시스템에 위임할 것인지 여부와 방법에 대해 선택할 수 있어야 합니다.

17) **비 전복** : 고도로 발전된 AI 시스템의 통제를 통해 부여되는 권력은 건강한 사회가 의존하는 사회적 시민적 과정을 전복하기 보다, 존중하고 개선해야 한다.

18) **AI 무기 경쟁** : 치명적인 자동화 무기의 군비 경쟁은 피해야 합니다.

## 장기적 이슈 (Longer-term Issues)

19) **능력치에 대한 주의** : 합의가 없으므로, 미래 AI의 능력 상한선에 대한 강한 가정은 피해야 합니다.

20) **중요성** : 고급 AI는  지구 생명체의 역사에서 중대한 변화를 나타낼 수 있으며, 그에 상응하는 관심 및 자원을 통해 계획되고 관리되어야 합니다.

21) **위험 요소** : AI 시스템이 초래하는 위험, 특히 치명적인 또는 실존적 위험은 예상되는 영향에 상응하여 대비하고 완화 노력을 기울여야 합니다.

22) **재귀적 자기 개선 : 질과 양을  빠르게 증가시킬 수 있도록 스스로  개선 또는 복제할 수 있도록 설계된 AI 시스템은 엄격한 안전 및 통제 조치를 받아야 합니다.**

23) **공동 선** : 수퍼 인텔리전스는 광범위하게 공유되는 윤리적 이상에만  복무하도록, 그리고 한 국가 또는 조직보다는 모든 인류의 이익을 위해 개발되어야 합니다.

## □ 2017년, EU 의회 결의안

ㅇ로봇 제작자들은 프로그램 오류나 외부에서의 해킹 등 비상 상황에서 로봇을 즉시 멈출 수 있는 'kill 스위치'를 장착해야 한다.

ㅇAI 로봇의 법적 지위를 전자적 인격체(electronic pesronhood)로 인정한다.

ㅇEU 내에 로봇담당국을 신설해, AI로봇에 대한 법적·윤리적 제재를 가할 수 있는 내용을 포함한다.
ㅇAI 로봇의 사회적 악용 가능성과 해킹에 대한 사회적 불안감을 해소하고 로봇이 인간에 무조건 복종하면서 위협을 가하지 않아야 한다.

ㅇAI 로봇 활용에 따른 새로운 고용 모델의 필요성을 강조하고, AI 로봇을 고용하는 기업에 '로봇세'를 물려야 한다.

https://brunch.co.kr/@kakao-it/184

## □ 2017년, 일본 총무성의 AI 개발 가이드라인

ㅇ(연계의 원칙) 개발자는 AI 시스템의 상호 접속성과 상호 운용성에 유의한다.

ㅇ(투명성의 원칙) 개발자는 AI 시스템 입출력의 검증 가능성과 판단 결과의 설명 가능성에 유의한다.

ㅇ(제어 가능성의 원칙) 개발자는 AI 시스템의 제어 가능성에 유의한다.

ㅇ(안전의 원칙) 개발자는 AI 시스템이 액추에이터 등을 통해 이용자와 제 3자의 생명·신체·재산에 위해를 미치는 것이 없게 배려한다.

ㅇ(보안의 원칙) 개발자는 AI 시스템의 안전에 유의한다.

ㅇ(프라이버시의 원칙) 개발자는 AI 시스템에 의해 이용자와 제 3자의 프라이버시가 침해되지 않도록 배려한다.

ㅇ(윤리의 원칙) 개발자는 AI 시스템 개발에 있어서 인간의 존엄과 개인의 자율을 존중한다.

ㅇ(이용자 지원의 원칙) 개발자는 AI 시스템이 이용자를 지원해 이용자에게 선택의 기회를 적절하게 제공하는 것이 가능하도록 배려한다.

ㅇ(책임의 원칙) 개발자는 이용자를 포함한 이해 관계자에 대해 책임을 완수하도록 노력한다.

https://brunch.co.kr/@kakao-it/184

# □ 2018.1.31 카카오 알고리즘 윤리헌장

**1. 카카오 알고리즘의 기본 원칙**

**카카오는 알고리즘과 관련된 모든 노력을 우리 사회 윤리 안에서 다하며, 이를 통해 인류의 편익과 행복을 추구한다.**

카카오가 알고리즘 윤리 헌장을 도입한 목적입니다.
카카오는 알고리즘 개발을 통해 카카오 서비스를 직·간접적으로 이용하는 사람들이 편익을 누리고, 보다 행복해지는 데 기여하고자 합니다. 알고리즘 개발 및 관리와 관련된 일련의 과정에서 카카오의 노력은 우리 사회의 윤리 원칙에 부합하는 방향으로 이뤄질 것입니다.

**2. 차별에 대한 경계**

**알고리즘 결과에서 의도적인 사회적 차별이 일어나지 않도록 경계한다.**

카카오는 다양한 가치가 공존하는 사회를 지향합니다.
카카오의 서비스로 구현된 알고리즘 결과가 특정 가치에 편향되거나 사회적인 차별을 강화하지 않도록 노력하겠습니다.

https://www.kakaocorp.com/kakao/ai/algorithm

### 3. 학습 데이터 운영

**알고리즘에 입력되는 학습 데이터를 사회 윤리에 근거하여 수집·분석·활용한다.**

카카오는 알고리즘의 개발 및 성능 고도화, 품질 유지를 위한 데이터 수집, 관리 및 활용 등 전 과정을 우리 사회의 윤리를 벗어나지 않는 범위에서 수행하겠습니다.

### 4. 알고리즘의 독립성

**알고리즘이 누군가에 의해 자의적으로 훼손되거나 영향받는 일이 없도록 엄정하게 관리한다.**

카카오는 알고리즘이 특정 의도의 영향을 받아 훼손되거나 왜곡될 가능성을 차단하고 있습니다.
앞으로도 카카오는 알고리즘을 독립적이고 엄정하게 관리할 것입니다.

### 5. 알고리즘에 대한 설명

**이용자와의 신뢰 관계를 위해 기업 경쟁력을 훼손하지 않는 범위 내에서 알고리즘에 대해 성실하게 설명한다.**

카카오는 새로운 연결을 통해 더 편리하고 즐거워진 세상을 꿈꿉니다.
카카오 서비스는 사람과 사람, 사람과 기술을 한층 가깝게 연결함으로써 그 목표에 다가가고자 합니다. 카카오는 모든 연결에서 이용자와의 신뢰 관계를 소중하게 생각합니다. 이를 위해 더 나은 가치를 지속적으로 제공하는 기업으로서, 이용자와 성실하게 소통하겠습니다.

FPRI
Future Policy Research Institute

## □ 2018.10.23 ICDPPC -인공지능 윤리 및 개인정보 보호에 대한 선언

* ICDPPC International Conference of Data Protection and Privacy Commissioners. 국제 개인정보보호기구회의. 세계 각국 개인정보보호 감독기구 대다수가 가입되어 있는 협의체로서, 정기 회의를 통해 개인정보 보호 관련 최신 의제와 공동의 입장을 논의함.

ICDPPC
International Conference of Data Protection & Privacy Commissioners

***1.*** 인공지능 및 머신러닝 기술은 기본적인 인권을 존중하며 설계되고 개발 및 이용되어야 하고, 다음과 같은 **공정성 원칙**을 따라야 한다.
a. 인공지능시스템을 본래 목적과 일치되게 사용하고 개인정보가 수집된 본래 목적에서 벗어나는 방식으로 사용되는 일이 없도록 보장함으로써 개인의 합리적인 기대 수준을 감안한다.
b. 인공지능의 사용이 개인에게 미치는 영향 뿐 아니라 집단이나 사회 일반에 미치는 집합적인 영향도 고려한다.
c. 인공지능 시스템이 인간 발전을 촉진하고 이를 방해하거나 위협하지 않는 방식으로 개발되도록 보장하고 특정한 용도에 대해서는 기술하고 한정할 필요가 있다.

https://act.jinbo.net/wp/40645/
https://icdppc.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf

***2.*** 인공지능시스템이 보장해야 할 책임성 뿐 아니라 그것이 미치는 잠재적인 영향 및 결과에 대해 다음과 같이 **지속적으로 주의하고 경계한다.**

a. 인공지능 시스템에 대한 사후감사, 지속적인 모니터링 및 영향평가를 실시하고 감독 체제에 대한 정기적인 검토 등을 통해 개인, 감독기구 및 기타 제3자에 이르기까지 모든 관련 이해관계자들의 책임성을 촉진한다.

b. 행위자 및 이해관계자의 전 단계에 협력 기준 개발과 모범사례 공유 등을 포함하여 집합적, 공동의 책임을 도모한다.

c. 인공지능에 대해 상당한 수준의 정보를 갖추기 위해 인식을 제고하고 교육하고 연구하고 훈련시키고, 인공지능과 그것이 사회적으로 미칠수 있는 영향을 이해할 수 있도록 노력한다.

d. 신뢰할 수 있는 제3자(TTP) 마련이나 독립적인 윤리위원회 설립 등 관련된 모든 행위자들에 대해 가시적인 거버넌스 절차를 수립한다.

***3.*** 인공지능 **시스템의 투명성 및 명료성(intelligibility)**이 향상되어야 하며, 다음과 같이 그 효과적인 구현을 목표로 한다.

a. 설명 가능한 인공지능(explainable artificial intelligence)에 대한 공공과 민간의 학술연구에 투자한다.

b. 혁신적인 통신 방식의 개발 등을 통해 투명성, 명료성, 접근가능성을 향상시키고 각각의 관련 당사자에게 필요한 다양한 수준의 투명성 및 정보를 고려한다.

c. 특히 알고리즘 투명성 및 시스템 감사가능성을 향상시켜 제공되는 정보의 의미성을 보장하는 한편, 기관의 관행을 보다 투명화한다.

d. 개인이 인공지능시스템과 직접 상호작용하거나 이들 시스템에 개인정보를 제공하여 처리되도록 할 때 언제든지 적절히 정보를 제공받을 수 있도록 보장하여 개인정보에 대한 자기결정권을 보장한다.

e. 시스템이 개인의 기대에 계속 일치하는지 확인하고 전반적으로 인간적으로 통제할 수 있도록 인공지능 시스템의 목적과 효과에 대한 적절한 정보를 제공한다.

**4. 전반적인 "윤리적 설계(ethics by design)" 접근의 일환으로서, 인공지능 시스템은 다음과 같이 프라이버시기본설정(privacy by default)과 프라이버시중심설계(privacy by design) 원칙을 적용함으로써 책임감 있게 설계되고 개발되어야 한다.**

a. 정보주체의 프라이버시권과 개인정보에 대한 권리를 보장하기 위해, 정보처리 수단을 결정할 때나 정보를 처리하는 순간 모두에서 개발 중인 시스템의 유형에 비례적인 기술 및 관리적 조치 및 절차를 실행한다.
b. 인공지능 프로젝트의 시작 단계 및 관련 개발의 모든 단계에서 개인과 사회에 미칠 것으로 예상되는 영향에 대해 평가하고 문서화한다.
c. 모든 인공지능 시스템의 개발과 운영의 일환으로 윤리적이고 공정한 시스템 이용과 인권 보장을 위한 요건을 구체적으로 명시한다.

5. 다음과 같이 모든 개인의 역량을 향상해야 하고, 개인의 권리 실현이 증진되어야 할 뿐 아니라, 대중 참여의 기회가 창출되어야 한다.

a. 개인정보에 대한 정보공개권, 열람권, 처리반대권, 삭제권 등 개인정보 및 사생활에 대한 권리를 적절하게 보장하고 교육 및 인식 제고 캠페인을 통해 이러한 권리들을 향상시킨다.
b. 차별금지, 표현 및 정보의 자유 등 관련된 권리를 보장한다.
c. 개인의 발전 및 의견에 영향을 미치는 기술들에 대해 반대할 권리 및 이의를 제기할수 있는 권리를 확인하고, 자동화된 처리가 개인의 권리에 상당히 영향을 미칠 때 오로지 여기에 기반하는 의사결정에 종속되지 않도록 개인의 권리를 적절히 보장하고, 부적절할 때는 그러한 의사결정에 문제를 제기할 수 있는 권리를 보장해야 한다.
d. 적응형 인터페이스 및 접근도구 등을 통해 개인 역량을 동등하게 향상시키고 대중적인 참여를 확대하기 위해 인공지능 시스템의 성능을 이용한다.

***6.*** 인공지능에서 개인정보의 사용으로 초래될 수 있는 불법적인 편향과 차별은 다음과 같이 감소되고 완화되어야 한다.

a. 인권 및 차별금지에 대한 국제법적 규범을 준수한다.
b. 편향을 식별하고 조치하고 완화할 수 있는 기술적 방식에 대한 연구에 투자한다.
c. 자동화된 의사결정에서 사용되는 개인정보가 정확하고 최신이며 가능한한 완전하도록 보장하는 합리적인 조치를 취한다.
d. 편향과 차별을 다루는 구체적인 지침과 원칙을 정교화하고 이 문제에 대한 개인 및 이해당사자의 인식을 제고한다.

FPRI
Future Policy Research Institute

## □ 2019.4.8 인공지능 윤리 가이드라인(Ethics guidelines for trustworthy AI)
## * 유럽연합(EU) 집행위원회(EC: European Commission) – AI-HLEG

EC의 독립된 AI 전문가 그룹(AI high Level Expert Group)이 '18.12월까지 500건 이상의 관련 의견을 수렴해 초안을 제시한데 이어 이번에 최종 버전을 완성

① 모든 시민이 AI 혜택을 누릴 수 있는 **인간 중심**의 윤리적 목적을 달성하는 동시에 **신뢰**할 수 있는 기술발전 기준을 구체적으로 제시

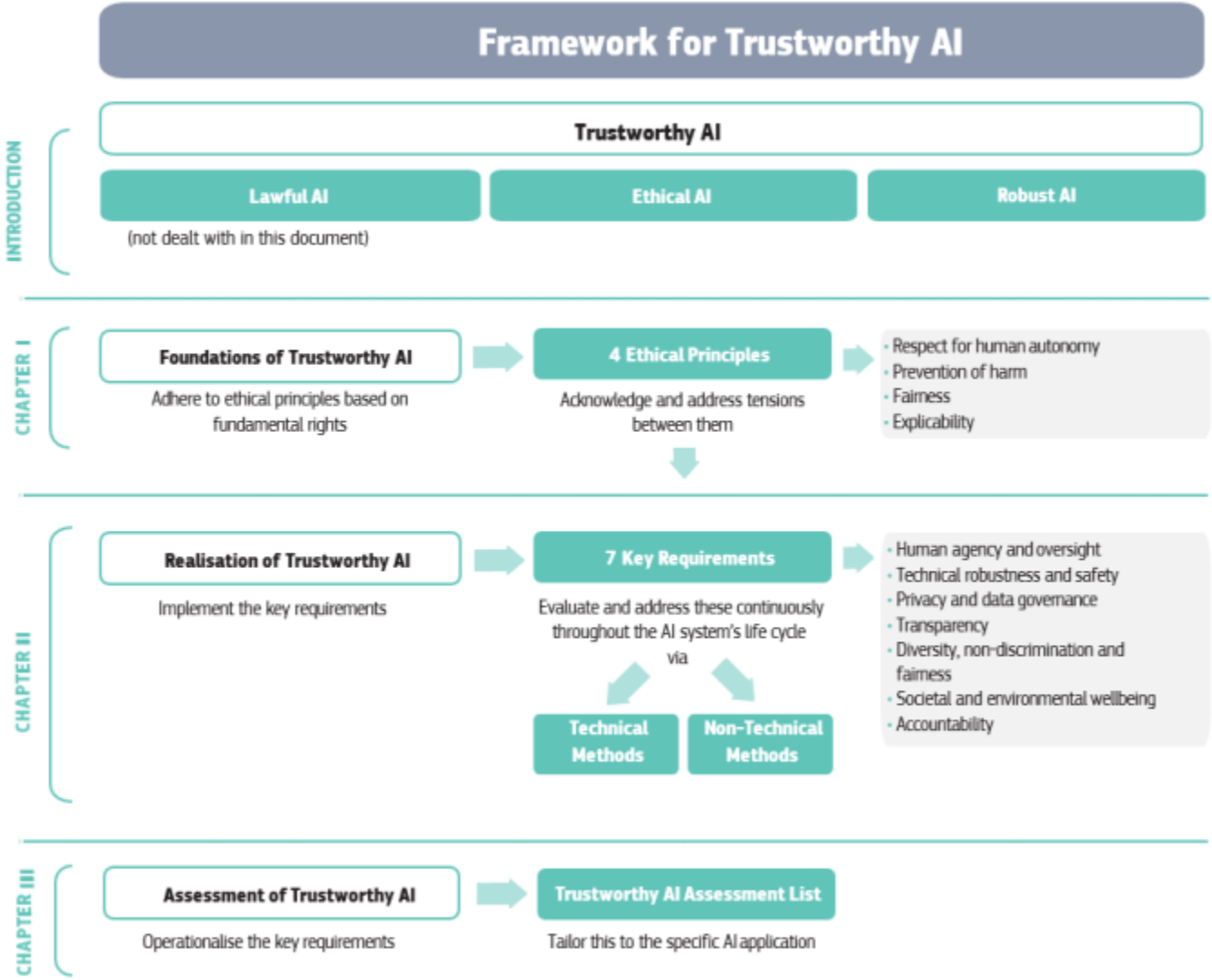
② 신뢰할만한 AI는 일련의 요구 사항뿐 아니라 관련된 모든 법률과 규정을 준수해야 하며 이를 검증하기 위해 특정 평가 목록이 수반되어야 한다고 설명
신뢰할 수 있는 AI를 확립하기 위해 3가지 요소와 각 분야별 핵심 지침을 제시

③ 개인과 사회에 막대한 이익을 창출하고 나아가 인류 번영과 성장을 촉진하는 AI의 가능성을 극대화하기 위해 적절한 관리와 안전한 법적 기틀 마련에 초점

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

FPRI
Future Policy Research Institute

# EU T-AI 가이드라인 : 4대 윤리원칙 → 7대 주요 요구사항 → 평가기준



FPRI
Future Policy Research Institute

| 3대 요소 | 핵심 지침 |
|---|---|
| I. Foundations of Trustworthy AI | ① **인간 존중을 윤리 원칙으로 준수하는 AI 시스템 개발·배포·사용**<br>- 자율성, 위해예방, 공정성 등을 고려<br>② **어린이·장애인·고용주와 근로자 또는 기업과 소비자 간에 권력이나 정보의 불균형에 대응**<br>- AI 기술이 불이익을 주거나 기술 혜택으로부터 배제 가능성이 있는 취약한 집단 배려<br>③ **AI기술이 개인과 사회에 상당한 혜택과 이익을 주지만 특정 위험도 초래할 가능성에 주의**<br>- 위험 강도에 따라 이를 완화하기 위한 적절한 조치 필요 |
| II. Realising Trustworthy AI | ① **인간의 기본권·존엄성·자율성 보장**<br>- AI 시스템은 인간의 기본권을 보장하고 자율성을 저해하지 않는 평등한 사회를 구현할 것<br>② **기술적 견고성 및 안전성**<br>- AI 시스템 알고리즘은 모든 라이프사이클 단계에서 오류와 오작동 등 처리가 가능한 안전성을 갖출 것<br>③ **개인정보 및 데이터거버넌스**<br>- 시민(개인)은 자신의 데이터(개인정보)를 완전히 삭제할 수 있어야 하며 관련 데이터가 인간에게 해를 입히거나 차별해서는 안 될 것<br>④ **투명성**<br>- AI 시스템은 설명 가능할 것<br>⑤ **다양성, 차별 금지 및 공정성**<br>- AI 시스템은 모든 범위의 인간 능력과 기술 및 요구 사항을 고려하고 접근성을 보장할 것<br>⑥ **사회·환경복지**<br>- AI 시스템은 긍정적인 사회 변화를 주도하고 지속 가능한 성장을 이끄는데 활용될 것<br>⑦ **책임성**<br>- AI 시스템과 그 결과에 대한 책임, 그 책임을 보장하기 위한 구조적 장치를 마련할 것 |
| III. Assessing Trustworthy AI | ① **II단계에서 언급한 요구 사항을 실제 사례에 적합하게 적용할 수 있는 기틀 마련**<br>- AI 시스템에 대한 요구 사항과 솔루션 평가 기준 확립<br>- AI 시스템의 라이프사이클 전반에 걸쳐 성과를 개선하고 이에 대한 이해 관계자 참여 등 |

https://now.k2base.re.kr/portal/trend/mainTrend/view.do?poliTrndId=TRND0000000000036251&menuNo=200043&pageIndex=1

FPRI
Future Policy Research Institute

# 7대 주요 요구사항 → (2)기술적 안전

## 2. Technical robustness and safety

***Resilience to attack and security:***

- ✓ Did you assess potential forms of attacks to which the AI system could be vulnerable?
  - Did you consider different types and natures of vulnerabilities, such as data pollution, physical infrastructure, cyber-attacks?
- ✓ Did you put measures or systems in place to ensure the integrity and resilience of the AI system against potential attacks?
- ✓ Did you verify how your system behaves in unexpected situations and environments?
- ✓ Did you consider to what degree your system could be dual-use? If so, did you take suitable preventative measures against this case (including for instance not publishing the research or deploying the system)?

***Fallback plan and general safety:***

- ✓ Did you ensure that your system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations (for example technical switching procedures or asking for a human operator before proceeding)?
- ✓ Did you consider the level of risk raised by the AI system in this specific use case?
  - Did you put any process in place to measure and assess risks and safety?
  - Did you provide the necessary information in case of a risk for human physical integrity?
  - Did you consider an insurance policy to deal with potential damage from the AI system?
  - Did you identify potential safety risks of (other) foreseeable uses of the technology, including accidental or malicious misuse? Is there a plan to mitigate or manage these risks?
- ✓ Did you assess whether there is a probable chance that the AI system may cause damage or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?
  - Did you consider the liability and consumer protection rules, and take them into account?
  - Did you consider the potential impact or safety risk to the environment or to animals?
  - Did your risk analysis include whether security or network problems such as cybersecurity hazards could pose safety risks or damage due to unintentional behaviour of the AI system?
- ✓ Did you estimate the likely impact of a failure of your AI system when it provides wrong results, becomes unavailable, or provides societally unacceptable results (for example discrimination)?
  - Did you define thresholds and did you put governance procedures in place to trigger alternative/fallback plans?
  - Did you define and test fallback plans?

***Accuracy***

- ✓ Did you assess what level and definition of accuracy would be required in the context of the AI system and use case?
  - Did you assess how accuracy is measured and assured?
  - Did you put in place measures to ensure that the data used is comprehensive and up to date?
  - Did you put in place measures in place to assess whether there is a need for additional data, for example to improve accuracy or to eliminate bias?
- ✓ Did you verify what harm would be caused if the AI system makes inaccurate predictions?
- ✓ Did you put in place ways to measure whether your system is making an unacceptable amount of inaccurate predictions?
- ✓ Did you put in place a series of steps to increase the system's accuracy?

***Reliability and reproducibility:***

- ✓ Did you put in place a strategy to monitor and test if the AI system is meeting the goals, purposes and intended applications?
  - Did you test whether specific contexts or particular conditions need to be taken into account to ensure reproducibility?
  - Did you put in place verification methods to measure and ensure different aspects of the system's reliability and reproducibility?
  - Did you put in place processes to describe when an AI system fails in certain types of settings?
  - Did you clearly document and operationalise these processes for the testing and verification of the reliability of AI systems?
  - Did you establish mechanisms of communication to assure (end-)users of the system's reliability?

FPRI
Future Policy Research Institute

# 7대 주요 요구사항 → (3) Privacy and data governance

### 3. Privacy and data governance

***Respect for privacy and data Protection:***

- ✓ Depending on the use case, did you establish a mechanism allowing others to flag issues related to privacy or data protection in the AI system's processes of data collection (for training and operation and data processing?

데이터의 수집, 처리시 프라이버시,정보보호

- ✓ Did you assess the type and scope of data in your data sets (for example whether they contain personal data)?
- ✓ Did you consider ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?

민감정보 & 개인정보의 최소 이용

- ✓ Did you build in mechanisms for notice and control over personal data depending on the use case (such as valid consent and possibility to revoke, when applicable)?
- ✓ Did you take measures to enhance privacy, such as via encryption, anonymisation and aggregation?

프라이버시 보호처리 – 암호화, 비식별화 등

- ✓ Where a Data Privacy Officer (DPO) exists, did you involve this person at an early stage in the process?

DPO의 참여

***Quality and integrity of data:***

- ✓ Did you align your system with relevant standards (for example ISO, IEEE) or widely adopted protocols for daily data management and governance?
- ✓ Did you establish oversight mechanisms for data collection, storage, processing and use?
- ✓ Did you assess the extent to which you are in control of the quality of the external data sources used?
- ✓ Did you put in place processes to ensure the quality and integrity of your data? Did you consider other processes? How are you verifying that your data sets have not been compromised or hacked?

***Access to data:***

- ✓ What protocols, processes and procedures did you follow to manage and ensure proper data governance?
  - Did you assess who can access users' data, and under what circumstances?
  - Did you ensure that these persons are qualified and required to access the data, and that they have the necessary competences to understand the details of data protection policy?
  - Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?

로깅을 통한 감독

## TODO

김명주 (2017),「인공지능 윤리의 필요성과 국내외 동향」,『한국통신학회지』34(10): 45-54.

**“만약 당신이 미래를 꿈꾸지 않거나 지금 기술개선을 위해 노력하지 않는다면 그건 곧 낙오되고 있는 것이나 마찬가지 입니다.”**

그윈 쇼트웰(Gwynne Shtwell, SpaceX CEO, COO)

# 감사합니다

(facebook.com/sangshik, mikado22001@yahoo.co.kr)