제19차 4차산업혁명위원회 심의안건 제2호
‘20. 12. 23. (수)

사람이 중심이 되는
「인공지능(AI) 윤리기준」

2020. 12. 23.

관계부처 합동

순 서

요 약

사람이 중심이 되는 「인공지능(AI) 윤리기준」 [요약]

1. 추진배경

□ 인공지능 기술의 급속한 발전으로 **제조·의료·교통·환경·교육 등 산업 全 분야**에 **인공지능이 활용·확산**되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 **인공지능 윤리 이슈가 새롭게 대두**

< 인공지능 윤리 이슈 사례 >

- **(기술오남용)** 유럽 한 에너지기업의 CEO는 영국 범죄자들이 AI를 활용해 정교하게 만든 모회사 CEO의 가짜음성에 속아 22만 유로를 송금하는 피해('19.9월)
- **(데이터 편향성)** 아마존의 인공지능 기반 채용시스템이 개발자, 기술 직군에 대부분 남성만을 추천하는 문제가 발생함에 따라 아마존에서 동 시스템 사용 폐기('18.10월)
- **(알고리즘 차별)** 인공지능 기반 범죄 예측 프로그램인 'COMPAS'의 재범률 예측에서 흑인 범죄자의 재범가능성을 백인보다 2배이상 높게 예측하는 편향 발견('18.1월)
- **(프라이버시 침해)** 아마존 '알렉사', 구글 '구글 어시스턴트', 애플 '시리' 등이 인공지능 스피커로 수집된 음성 정보를 제3의 외부업체가 청취하는 것으로 밝혀져 논란(UPI, '19.9월)

□ OECD, EU 등 세계 각국과 주요 국제기구 등은 인공지능 윤리의 중요성을 인식하고 **윤리적인 인공지능 실현을 위한 원칙**들을 **발표 중**

< 인공지능 윤리 규범 동향 >

구분	윤리 규범
OECD	인공지능 권고안('19.5, 이사회), 동 권고안의 G20 정상선언문 반영('19.6)
EU	신뢰할 수 있는 인공지능 윤리 가이드라인('18.12, 인공지능 고위전문가그룹)
UNESCO	인공지능 윤리에 대한 권고사항 초안('19.5, 특별전문가그룹)
일본	인간 중심의 인공지능사회 원칙('19.3, 통합혁신전략추진회의)
미국	AI 활용에 대한 구글 원칙('18.6, 구글)

○ **우리나라**에서도 '18.4월 **「지능정보사회윤리가이드라인①(정보문화포럼)」**, **'19.11월 「이용자중심의 지능정보사회를 위한 원칙②(방통위)」** 등이 마련된 바 있으나, **범국가 인공지능 윤리원칙이라기**에는 **다소 제한적인 측면**

* 두 원칙 모두 AI 특화가 아닌 지능정보기술 전반(클라우드, 빅데이터, IoT, 블록체인 등)을 대상으로 하며, ①은 정부가 아닌 민간 발표, ②는 이용자 보호 관점을 강조

⇨ **「인공지능 국가전략」**에 따른 **'사람 중심의 인공지능' 구현**을 위해 **글로벌 기준과의 정합성**을 갖춘 **'인공지능 윤리기준'** 마련 추진

사람이 중심이 되는 인공지능 윤리기준의 목표 및 지향점

❶ 모든 사람이 ❷ 모든 분야에서 ❸ 자율적으로 준수하며 ❹ 지속 발전

❶ 인공지능 전 주기에서 모든 사회 구성원이 참조할 수 있는 기준

○ '사람중심의 인공지능' 구현을 위해 **인공지능의 개발부터 활용에 이르는 전 과정**에서 **정부·공공기관, 기업, 이용자 등 모든 사회 구성원이 참조**할 수 있는 기준

❷ 특정 분야에 제한되지 않는 범용성을 가진 일반 원칙

○ 인공지능과 관련한 전 분야에 대한 참조모델이 되는 **총론 차원의 윤리 기준**을 제시하고, **사안별 또는 분야별 인공지능 윤리기준 제정의 근거**를 제공

❸ '법'이나 '지침'이 아닌 자율 규범

○ 구속력 있는 '법'이나 '지침'과 별개로 '윤리 기준'을 제시함으로써 **기업 자율성을 존중하고 기술발전을 장려**하며, **기술·사회변화에 유연하게 대처**할 수 있는 기반 마련

❹ 새롭게 제기되는 인공지능 윤리 이슈를 논의·발전시키는 플랫폼

○ **사회경제, 기술적 변화**와 함께 개별 영역에서 **새롭게 제기되는 윤리적 이슈를 논의하고 구체적으로 발전**시킬 수 있는 **플랫폼**으로 기능

2. 그간의 추진경과

1 윤리기준 초안 마련('20.4~8월)

○ (**윤리연구반 구성·운영**) 인공지능·윤리 전문가로 구성된 인공지능 윤리 연구반을 통해 주요국, 국제기구, 학회, 기업 등에서 발표된 **국내외 주요 인공지능 윤리원칙(25개)**을 **주체·목적·특징별**로 비교·분석 수행

국내외 주요 인공지능 윤리원칙 사례

- (**OECD**) OECD '디지털경제정책위원회' 주관하에 신뢰가능한 AI를 위한 5개 원칙 및 5개 제언을 담은 「Recommendation of the Council on AI」 발표('19.5)
- (**EU**) EU 산하 AI고위전문가 그룹 주도로 관련 주체들에게 필요한 7개 윤리 원칙과 각 원칙별 평가리스트를 담은 「Ethics Guideline for Trustworthy AI」 발표('18.12)
- (**일본**) 일본의 AI 연구개발 목표 및 산업화 로드맵에 따라 25명의 산학연 전문가로 구성된 '인간 중심의 AI 사회 원칙 위원회'를 통해 「인간 중심의 AI 원칙」 발표('19.3)

○ (**윤리기준 체계 구조화**) 주요 인공지능 **윤리 원칙 분석 결과**를 토대로 '인간성을 위한 인공지능(AI for Humanity)'를 구성하는 **4개 속성(四端)**과 이를 달성하기 위한 **3대 기본원칙(三綱), 15대 실행원칙(十五倫)**을 **구조화**

* **4단(四端)**: 인공지능이 인간성 실현을 위해 근본적으로 지니고 있어야 할 속성
* **3강(三綱)**: 인간과 인공지능의 관계에서 인공지능이 마땅히 준수해야할 기본원칙
* **15륜(十五倫)**: 기본원칙(三綱)을 실행하기 위한 세부적인 실행원칙

2 각계 의견수렴('20.9~12월)

○ **(초안 의견수렴) 학계·기업·시민단체**를 아우르는 **각계 전문가***를 대상으로 **윤리기준 초안에 대한 폭넓은 의견수렴(9~11월)**

* (**학계**) 전자전기공학, 컴퓨터공학, 법학, 인문사회학, 윤리학, 의료법윤리학 등
* (**기업**) 네이버, 카카오, 삼성전자, LG전자, IBM, 통신3사, 현대차, 기업 협회·단체 등
* (**시민단체**) 소비자시민모임, 진보네트워크센터, 오픈넷, 녹색소비자연대 등

각계 주요의견

- (**공통의견**) **전체적으로 간결화·최소한의 원칙만 제시**하는 것이 바람직, **4단과 3강이 중복**되며 '기술윤리적 좋음' 등 **일부 용어는 이해가 어려움**
- (**학계**) ① 기술의 불확실성 등 한계를 고려하여 **견고성 원칙 삭제** ② 15대 원칙 간 **최대한 중복이 없도록 조정** ③ 각 원칙간 **상충관계 고려** 필요 ④ **규제로 오인되지 않도록 문구 수정**
- (**기업**) ① 기술의 한계 고려 **견고성 원칙은 삭제 필요** ② 인공지능이 **독립적 인격이 아니며 동 기준이 법규가 아님을 명확히** 할 필요 ④ 현장에서 참고할 **구체적 가이드 필요**
- (**시민단체**) ① 강(綱)·륜(倫)과 같은 **유교적 표현은 젊은 세대·일반 대중이 느끼기에 다소 난해** ② **투명성 원칙 내용을 강화할 필요**

⇒ 각계 의견을 반영하여 구조를 간결화하고 강(綱)·륜(倫)과 같은 표현을 수정하는 등 보완, **3대원칙·10대요건으로 구성된 최종안 마련**

○ **(최종안 의견수렴) 공개 공청회 개최***(12.7) 및 인공지능 **기술 전문가 그룹·시민 의견 접수**(11.25~12.15) 등 윤리기준 최종안에 대한 **공개 의견수렴 실시(12월)**

* 추진 경과·주요내용 소개, 각계 전문가(11명) 토론, 현장·온라인 질의 응답 등 진행

각계 주요 의견

- (**윤리기준 필요성 공감**) 인공지능 윤리이슈를 논의하는 시작점이자 플랫폼으로서 의미
- (**과도한 불안감 조성·규제 지양**) 인공지능에 대한 과도한 불안감 조성이나 기술 발전을 저해할 수 있는 규제로 이어지는 것은 경계할 필요
- (**기술을 통한 문제해결**) 알고리즘 차별, 데이터편향성 등 해결을 위한 연구개발 투자 중요
- (**인공지능 교육 필요**) 인공지능 기술과 윤리 모두에 대한 국민 교육이 필요
- (**폭넓은 의견수렴**) 지속적으로 각계 전문가 의견을 수렴하고 반영해온 과정이 모범적

◇ 윤리기준(안) 자체 보다는 **활용 및 실천에 관련된 제언이 다수**인만큼, 향후 주체별 체크리스트·윤리 교육 등 **실천방안 마련을 차질 없이 추진**

3. 사람이 중심이 되는 「인공지능(AI) 윤리기준」 주요내용

□ (**서문**) 윤리기준 마련의 시대적 배경과 **윤리기준이 지향하는 목표 및 지향점, '인간성을 위한 인공지능' 가치** 제시, **상충관계 조항** 등 설명

□ (**3대 기본원칙**) '인간성을 위한 인공지능'을 위해 고려해야할 기본원칙

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

② 사회의 공공선 원칙
- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙
- 인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.
- 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.

□ (**10대 핵심요건**) 3대 기본원칙을 실현하기 위한 세부적인 요건

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·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 직·간접적인 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 인공지능의 투명성과 설명 가능성을 높이고, 타 원칙과의 상충관계를 고려하여 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

□ **(부록)** 「인공지능 윤리기준」에서 인공지능의 지위, 윤리기준의 적용범위와 대상, 윤리기준의 실현방안 등 제시

4. 향후 계획

□ (**주체별 체크리스트 마련**) 윤리기준의 활용도 제고 및 사회 확산을 돕기 위해 AI 개발~활용 전 과정에 참여하는 **주체별* 윤리 점검 체크리스트 개발 및 보급**

* (예시) ① 개발자(데이터 확보 및 AI 설계시), ② 제공자(AI기반 서비스·제품 제공시), ③ 이용자(AI 제품 구매·활용시), ④ 정부·공공기관(정책개발·집행) 등

□ (**교육 프로그램 마련**) 전문인력, 일반시민, 개발자 등 다양한 사회구성원으로 구분하여 생애단계별 **인공지능 윤리 교육 커리큘럼 연구·개발**

□ (**후속 보완 등**) 동 윤리기준을 기본 플랫폼으로 새로운 인공지능 윤리 이슈 지속 논의, **필요시 윤리기준 보완 및 세부기준 마련·입법 지원**

본 문

Ⅰ. 추진배경

□ 인공지능 기술의 사회·산업 확산 → 인공지능 윤리 이슈 대두

○ 인공지능 기술의 급속한 발전으로 **제조·의료·교통·환경·교육 등 산업 全 분야에 인공지능이 활용·확산**되면서, 기술의 오남용·알고리즘에 의한 차별·프라이버시 침해 등 **인공지능 윤리 이슈가 새롭게 대두**

< 인공지능 윤리 이슈 사례 >

- **(기술오남용)** 유럽 한 에너지기업의 CEO는 영국 범죄자들이 AI를 활용해 정교하게 만든 모회사 CEO의 가짜음성에 속아 22만 유로를 송금하는 피해('19.9)
- **(데이터 편향성)** 아마존의 인공지능 기반 채용시스템이 개발자, 기술 직군에 대부분 남성만을 추천하는 문제가 발생함에 따라 아마존에서 동 시스템 사용 폐기('18.10월)
- **(알고리즘 차별)** 인공지능 기반 범죄 예측 프로그램인 'COMPAS'의 재범률 예측에서 흑인 범죄자의 재범가능성을 백인보다 2배이상 높게 예측하는 편향 발견('18.1월)
- **(프라이버시 침해)** 아마존 '알렉사', 구글 '구글 어시스턴트', 애플 '시리' 등이 인공지능 스피커로 수집된 음성 정보를 제3의 외부업체가 청취하는 것으로 밝혀져 논란(UPI, '19.9)

□ 국제사회 인공지능 윤리원칙의 등장과 필요성

○ OECD, EU 등 **세계 각국과 주요 국제기구 등**은 인공지능 윤리의 중요성을 인식하고 **윤리적인 인공지능 실현을 위한 원칙**들을 **발표 중**

< 인공지능 윤리 규범 동향 >

구분	윤리 규범
OECD	인공지능 권고안('19.5, 이사회), 동 권고안의 G20 정상선언문 반영('19.6)
EU	신뢰할 수 있는 인공지능 윤리 가이드라인('18.12, 인공지능 고위전문가그룹)
UNESCO	인공지능 윤리에 대한 권고사항 초안('19.5, 특별전문가그룹)
일본	인간 중심의 인공지능사회 원칙('19.3, 통합혁신전략추진회의)
미국	AI 활용에 대한 구글 원칙('18.6, 구글)

○ **우리나라**에서도 '18.4월 **「지능정보사회윤리가이드라인①(정보문화포럼)」**, **'19.11월 「이용자중심의 지능정보사회를 위한 원칙②(방통위)」** 등이 마련된 바 있으나, **범국가 인공지능 윤리원칙이라기**에는 **다소 제한적인 측면**

* 두 원칙 모두 AI 특화가 아닌 지능정보기술 전반(클라우드, 빅데이터, IoT, 블록체인 등)을 대상으로 하며, ①은 정부가 아닌 민간 발표, ②는 이용자 보호 관점을 강조

□ 글로벌 기준에 부합하는 한국형 '인공지능 윤리기준' 마련

○ 윤리기준은 **관련 규범이 불확실하거나 법·제도가 급속한 기술발전을 따라가지 못할 경우**에도 **우리사회가 나아가야 할 방향**을 **제시**하는 의의

⇨ 이에 **「인공지능 국가전략」**에 따른 **'사람 중심의 인공지능'** 구현을 위해 **글로벌 기준과의 정합성**을 갖춘 **'인공지능 윤리기준'** 마련 추진

인공지능 윤리기준의 목표 및 지향점

❶ 모든 사람이 ❷ 모든 분야에서 ❸ 자율적으로 준수하며 ❹ 지속 발전

❶ 인공지능 전 주기에서 모든 사회 구성원이 참조할 수 있는 기준

○ '사람중심의 인공지능' 구현을 위해 **인공지능의 개발부터 활용에 이르는 전 과정**에서 **정부·공공기관, 기업, 이용자 등 모든 사회 구성원이 참조**할 수 있는 기준

❷ 특정 분야에 제한되지 않는 범용성을 가진 일반 원칙

○ **범용성을 가진 일반원칙**으로서의 윤리 기준을 제시하여 **각 영역별 세부 규범이 유연하게 발전해 나갈 수 있는 기반을 조성**

- 즉 인공지능과 관련한 전 분야에 대한 참조모델이 되는 **총론 차원의 윤리기준**을 제시하고, **사안별 또는 분야별 인공지능 윤리기준 제정의 근거**를 제공

❸ '법'이나 '지침'이 아닌 자율 규범

○ 구속력 있는 '법'이나 '지침'과 별개로 '윤리 기준'을 제시함으로써 **기업 자율성을 존중하고 기술발전을 장려**하며, **기술·사회변화에 유연하게 대처**할 수 있는 기반 마련

✓ 인공지능 윤리 기준은 **보편적 도덕가치**에 기반한 **포괄적인 원칙**이자 **강제력 없는 도덕 규범**

✓ 인공지능 개발 및 활용과 관련하여 **명백한 법적 규제가 필요한 개별 사안**들은 **사회적 합의를 토대로 관계법령에 개별적으로 규정 가능**

❹ 새롭게 제기되는 인공지능 윤리 이슈를 논의·발전시키는 플랫폼

○ **사회경제, 기술적 변화**와 함께 개별 영역에서 **새롭게 제기되는 윤리적 이슈를 논의하고 구체적으로 발전**시킬 수 있는 **플랫폼**으로 기능

Ⅱ. 그간의 추진 경과

1 윤리기준 초안 마련(20.4~8)

□ 인공지능 윤리연구반 구성·운영

○ 인공지능·윤리 전문가로 구성된 인공지능 윤리연구반을 통해 주요국, 국제기구, 학회, 기업 등에서 발표된 **국내외 주요 인공지능 윤리원칙(25개)**을 **주체·목적·특징별**로 비교·분석 수행('참고2')

국내외 주요 인공지능윤리원칙 현황

- **(OECD) OECD '디지털경제정책위원회'** 주관하에 신뢰가능한 AI를 위한 5개 원칙 및 5개 제언을 담은 **「Recommendation of the Council on AI」 발표('19.5)**
- **(주요 원칙)** 포용적 성장, 지속가능한 발전, 인간중심 가치, 공정성, 투명성, 설명가능성, 견고성, 보안 및 안전, 책무성
- **(EU)** EU 산하 **AI고위전문가 그룹** 주도로 관련 주체들에게 필요한 **7개 윤리 원칙과 각 원칙별 평가리스트**를 담은 **「Ethics Guideline for Trustworthy AI」 발표**('18.12)
- **(주요 원칙)** 인간행위자와 감독, 기술적 견실성(Technical robustness), 안전, 사생활, 데이터 관리, 투명성, 다양성, 차별금지, 정당성, 사회환경적 복지, 책무성 등
- **(일본)** 일본의 AI 연구개발 목표 및 산업화 로드맵에 따라 25명의 산학연 전문가로 구성된 **'인간 중심의 AI 사회 원칙 위원회'**를 통해 **「인간 중심의 AI 원칙」 발표**('19.3)
- **(주요 원칙)** 인간 중심 AI, 교육 평등 제공, 개인정보 보호, 보안 확보, 공정 경쟁, 공정·책임·투명성, 혁신

□ 인간성(Humanity)을 중심으로 인공지능 윤리체계 구조화

○ 국내외 주요 인공지능 **윤리 원칙 분석 결과**와 연계, **윤리철학 이론 토대로** '인공지능 윤리기준' 체계 구조화

⇨ '사람 중심의 인공지능 구현'을 위해 **윤리기준이 지향하는 최고가치를 '인간성(Humanity)**로 설정

⇨ '인간성을 위한 인공지능(AI for Humanity)'**를 구성하는 4개 속성**과 이를 달성하기 위한 **3대 기본원칙 15대 실행원칙**을 **수직적으로 구조화**

인공지능 윤리기준 초안 구조

□ **인공지능이 지향하는 최고의 가치를 '인간성(Humanity)'**으로 설정하고, 모든 인공지능은 **'인간성을 위한 인공지능(AI for Humanity)'**여야 함을 명시

□ '인간성을 위한 인공지능(AI for Humanity)'가 지녀야할 **4대 속성(四端)**을 설명하고, 이를 위해 준수해야 할 **3대 기본원칙(三綱) 및 15대 실행원칙(十五倫)** 제시

- **4단(四端)**: 인공지능이 인간성 실현을 위해 근본적으로 지니고 있어야 할 속성
- **3강(三綱)**: 인간과 인공지능의 관계에서 인공지능이 마땅히 준수해야할 기본원칙
- **15륜(十五倫)**: 기본원칙(三綱)을 실행하는 세부적인 기준

<전체 구조>

<15대 실행원칙 주요 키워드>

2 각계 의견 수렴(20.9~12)

◇ 「인공지능 윤리기준」 초안에 대한 각계 전문가 의견 수렴(9~11월)

□ 개 요

○ (전문가 의견수렴) AI·SW 자문위원회, AI 법제정비단 등 학계·기업·시민단체 전문가*를 대상으로 윤리기준 초안에 대한 폭넓은 의견수렴

* (**학계**) 전자전기공학, 컴퓨터공학, 법학, 인문사회학, 윤리학, 의료법윤리학 등
* (**기업**) 네이버, 카카오, 삼성전자, LG전자, IBM, 통신3사, 현대차, 기업 협회·단체 등
* (**시민단체**) 소비자시민모임, 진보네트워크센터, 오픈넷, 녹색소비자연대 등

□ 주체별 주요 의견 및 검토 결과 ('참고3')

○ (**공통의견**) **전체적으로 간결화·최소한의 원칙**만 제시하는 것이 바람직, **4단과 3강이 중복**되고 **'기술윤리적 좋음' 등** 일부 용어는 **이해가 어려움**

⇒ **4단을 삭제**하여 중복·혼란 해소, **15대 실행원칙을 10대 요건으로 조정 등 보완**

○ (**학계**) ① 인공지능 기술의 **불확실성 등 한계**를 **고려**하여 **견고성 원칙 삭제** ② **15대 원칙 간** 최대한 **중복이 없도록** 조정, ③ **각 원칙간 상충관계**에 대한 **고려** 필요 ④ **규제로 오인되지 않도록** 문구 수정

⇒ 지나치게 높은 수준의 요구라는 의견이 많은 **견고성원칙은 삭제, 15대 원칙을 10대 요건**으로 조정하고 윤리기준 **전문에** 원칙간 **상충관계 언급 등 보완**

○ (**기업**) ① 現 기술의 한계 감안, **견고성 원칙은 삭제 필요** ② **인공지능이 독립적 인격이 아님**을 분명히 할 필요, ③ **동 기준이 법규가 아님을 명확히 할 필요** ④ 현장에서 참고할 **구체적 가이드 필요**

⇒ **견고성 원칙은 삭제**하고 인공지능이 독립적 인격으로 오해되지 않도록 **문장 주술구조 등 수정**, 윤리기준 배포시 **주체별 체크리스트 개발 계획 안내 등**

○ (**시민단체**) ① **강(綱)·륜(倫)과 같은 유교적 표현**은 젊은 세대·일반 대중이 느끼기에 **다소 난해** ② **투명성 원칙 내용**을 **강화**할 필요

⇒ **단·강·륜 등 성리학적 표현은 삭제**, 다만 **투명성 원칙 강화**는 현재의 기술 수준을 고려해야 한다는 **학계·기업 다수의견과 충돌되어 미수용**

◇ 보완된 「인공지능 윤리기준」 최종안에 대한 공개 의견수렴(12월)

□ 개 요

○ **공개 공청회 개최***(12.7) 및 인공지능 **기술 전문가 그룹·시민 의견 접수**(11.25~12.15) 등 윤리기준 최종안에 대한 공개 의견수렴 실시

* 추진 경과·주요내용 소개, 각계 전문가(11명) 토론, 현장·온라인 질의 응답 등 진행

□ 주요 의견

① 전체 방향 관련

○ (**과도한 불안감 조성·규제 지양**) 인공지능에 대한 **과도한 불안감 조성이나 기술 발전을 저해할 수 있는 규제**로 이어지는 것은 **경계**할 필요

② 내용·수립과정 관련

○ (**기술수준 고려**) 현재의 기술수준을 고려해 **요구수준이 지나치게 높지 않은 원칙을 세운 것**은 **바람직하며 적절**한 조치

○ (**폭넓은 의견수렴**) 지속적으로 각계 전문가 의견을 수렴하고 윤리기준(안)에 반영해온 과정이 **진정성 있고 모범적**으로 평가

③ 실천방안·후속조치 관련

○ (**구체적 가이드 필요**) 인공지능 윤리기준을 **실제 현장에서 실천**할 수 있도록 **주체별 체크리스트 등 구체적인 가이드**도 **필요**

○ (**관련 연구개발 투자**) 기술을 통해서도 알고리즘 차별이나 데이터 편향성을 해결할 수 있는 만큼 **관련 연구개발 투자 지원 등 필요**

○ (**기술·윤리 교육**) 기술에 대한 국민들의 이해가 없다는 것이 큰 도전, **인공지능 기술과 윤리 모두에 대한 국민 교육**이 필요

○ (**신규 이슈·기술 반영**) 앞으로도 기업·시민단체 의견을 폭넓게 수렴하면서 윤리기준을 지속적으로 업데이트 할 필요

◇ 「인공지능 윤리기준」(안) 자체 보다는 **향후 활용 및 실천에 관련 제언 다수**
⇒ 향후 **주체별 체크리스트 개발, 윤리 교육 등** 실천방안 마련을 **차질 없이 이행**

Ⅲ. 사람이 중심이 되는 인공지능 윤리기준 최종본 주요내용(참고1)

□ 서문

○ 윤리기준 마련의 시대적 배경과 **윤리기준이 지향하는 목표 및 지향점, '인간성을 위한 인공지능'** 개념 제시, **상충관계 조항** 등 설명

서문 주요내용

- **(시대적 배경) 컴퓨팅 파워의 성장, 데이터의 축적, 네트워크 고도화**와 같은 ICT 기술의 발전을 토대로 **인공지능 기술 급성장 및 사회·산업 확산**
  - 국가경쟁력 제고 및 국민 삶의 질 개선 등 순기능과 함께 **데이터 편향성, 기술 오용과 같은 인공지능 윤리 이슈가 대두**
- **(목표 및 지향점) 인공지능 개발~활용 전 단계**에서 정부·공공기관, 기업, 이용자 등 **모든 사회구성원이 고려**해야할 **기본적·포괄적 기준 제시**
  - **산업·경제분야의 자율 규제 환경을 조성**함으로써 인공지능 연구개발과 산업 성장을 제약하지 않으며, **분야별 세부규범의 발전 근거**가 됨
- **(인간성을 위한 AI)** 인공지능은 **인간성(Humanity)를 최고가치로 지향**해야 하며, **인간에게 유용**하고 **정신과 신체에 해롭지 않도록** 개발·활용되어야 하며, **개인의 행복과 사회의 긍정적 변화 등에 기여**해야함

□ 3대 기본원칙

○ '인간성을 위한 인공지능(AI for Humanity)'을 위해 인공지능 개발에서 활용에 이르는 전 과정에서 고려해야할 3대 기본원칙을 제시

① 인간 존엄성 원칙

- 인간은 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계 제품과는 교환 불가능한 가치가 있다.
- 인공지능은 인간의 생명은 물론 정신적 및 신체적 건강에 해가 되지 않는 범위에서 개발 및 활용해야 한다.
- 인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.

② 사회의 공공선 원칙

- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙

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

□ 10대 핵심요건 - 기본원칙을 실현할 수 있는 세부적인 요건

o 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건을 제시

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·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 직·간접적인 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 인공지능의 투명성과 설명 가능성을 높이고, 타 원칙과의 상충관계를 고려하여 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

□ 부록

o 인공지능 윤리기준에서 인공지능의 지위, 윤리기준의 적용범위와 대상, 윤리기준의 실현방안 등 제시

Ⅵ. 향후계획

◇ 「인공지능 윤리기준」 실천방안 마련

□ 인공지능 생태계 참여 주체별 체크리스트 마련

o **윤리기준의 활용도 제고 및 사회 확산**을 돕기 위해 AI 개발~활용 전 과정에 참여하는 **주체별* 윤리 점검 체크리스트 개발 및 보급****

* (예시) ① 개발자(데이터 확보 및 AI 설계시), ② 제공자(AI기반 서비스·제품 제공시), ③ 이용자(AI 제품 구매·활용시), ④ 정부·공공기관(정책개발·집행) 등

** AI 윤리기준에서 제시하고 있는 3대 원칙·10대 요건 각각에 대해 주체별 점검 등

□ 인공지능 윤리 교육 프로그램 마련

o 전문인력, 일반시민, 개발자 등 다양한 사회구성원으로 구분하여 생애단계별 **인공지능 윤리 교육 커리큘럼 연구·개발**

* (예시) 알고리즘 기초 교육과 AI 사용 윤리 교육 방안 연구, AI의 윤리적 사용을 위한 시민 교육 및 지식 확산 프로그램 연구 등

◇ 인공지능 윤리 관련 논의·발전을 위한 장 마련

□ 인공지능 윤리 포럼 운영

o 인공지능 **윤리기준을 기본 플랫폼**으로 하여 학계·기업·시민 단체 등 다양한 이해관계자간 참여를 통해 새로운 **윤리 이슈와 쟁점을 토론·논의할 수 있는 '인공지능 윤리 포럼' 운영**

□ 범정부 인공지능 윤리 협력체계 구축

o **국방, 산업, 의료 등 분야별로 제기되는 인공지능 윤리 이슈**에 대해 **분야별 윤리기준·세부 지침 마련** 등 대응할 수 있도록 관계 부처 회의 등 범정부 협력체계 구축

사람이 중심이 되는 「인공지능(AI) 윤리기준」

2020. 12

관계부처 합동

Ⅰ. 서문

오늘날 인공지능 기술은 컴퓨팅 파워의 성장, 데이터의 축적, 5G 등 네트워크 고도화와 같은 ICT 기술의 발전을 토대로 급 성장하고 있다. 인공지능은 제조, 의료, 교통, 환경, 교육 등 산업 전반에서 본격적으로 활용·확산되고 있으며, 우리 생활에서도 쉽게 인공지능 기술을 접할 수 있게 되었다. 이러한 인공지능 기술의 발전·확산은 생산성·편의성을 높여 국가 경쟁력을 높이고 국민의 삶의 질을 높일 것으로 기대되지만, 한편으로는 기술 오용, 데이터 편향성과 같은 인공지능 윤리 이슈도 제기되고 있다. 본 윤리기준은 이러한 시대적 흐름을 고려하여 '인공지능 개발과 활용 전 단계에서 정부·공공기관, 인공지능 기술 개발자, 인공지능 기술을 활용한 제품·서비스 공급자·활용자 등 모든 사회 구성원이 사람중심의 인공지능' 구현을 위해 고려해야 할 기본적이고 포괄적인 기준을 제시하는 것을 목표로 한다.

본 윤리기준은 '사람중심의 인공지능' 구현을 위해 지향되어야 할 최고 가치로 '인간성(Humanity)'을 설정하고 있다. 이는 아래와 같은 사실을 의미한다. 모든 인공지능은 '인간성을 위한 인공지능(AI for Humanity)'을 지향하고, 인간에게 유용할 뿐만 아니라 나아가 인간 고유의 성품을 훼손하지 않고 보존하고 함양하도록 개발되고 활용되어야 한다. 인공지능은 인간의 정신과 신체에 해롭지 않도록 개발되고 활용되어야하며, 개인의 윤택한 삶과 행복에 이바지하며 사회를 긍정적으로 변화하도록 이끄는 방향으로 발전되어야 한다. 또한 인공지능은 사회적 불평등 해소에 기여하고 주어진 목적에 맞게 활용되어야 하며, 목적의 달성 과정 또한 윤리적이어야 하고, 궁극적으로 인간의 삶의 질 및 사회적 안녕과 공익 증진에 기여 하도록 개발되고 활용되어야 한다.

본 윤리기준은 산업·경제 분야의 자율규제 환경을 조성함으로써 인공지능 연구개발과 산업 성장을 제약하지 않고, 정당한 이윤을 추구하는 기업에 부당한 부담을 지우지 않는 것을 목표로 한다. 또한 본 윤리기준은 범용성이 있는 일반 원칙으로서 사안별 또는 분야별 인공지능 윤리기준 제정의 근거를 제공하여 영역별 세부 규범이 유연하게 발전해 나갈 수 있는 기반을 조성하고, 나아가 사회경제 및 기술 변화와 함께 새롭게 제기되는 인공지능 윤리 쟁점을 반영하여 지속적으로 수정되고 보완되는 일종의 '인공지능 윤리 플랫폼'으로 기능할 수 있다.

본 윤리기준에서 제시하는 원칙과 요건들은 상황에 따라 상충관계가 발생할 수 있으며, 상충하는 문제의 해결 방식은 개별 맥락과 상황에 따라 달라질 수 있다. 따라서 본 윤리기준에서는 각각 원칙들 사이에 고정된 형태의 우선순위를 제시하지는 않으며, 직간접적으로 영향을 받는 이해관계자가 지속적인 토론과 숙의 과정에 참여하여 절충점과 해결 방안을 모색하도록 권유한다.

Ⅱ. 인공지능 윤리기준: 3대 기본원칙, 10대 핵심요건

1. 3대 기본원칙 - 인공지능 개발 및 활용 과정에서 고려될 원칙

o **'인간성을 위한 인공지능(AI for Humanity)'**을 위해 **인공지능 개발에서 활용에 이르는 전 과정**에서 고려되어야할 기준으로 **3대 기본원칙**을 제시한다.

① 인간 존엄성 원칙

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

② 사회의 공공선 원칙

- 공동체로서 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록 개발 및 활용되어야 한다.
- 공익 증진을 위한 인공지능 개발 및 활용은 사회적, 국가적, 나아가 글로벌 관점에서 인류의 보편적 복지를 향상시킬 수 있어야 한다.

③ 기술의 합목적성 원칙

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

2. 10대 핵심요건 - 기본원칙을 실현할 수 있는 세부 요건

o 3대 기본원칙을 실천하고 이행할 수 있도록 **인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건**을 제시한다.

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.
- 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

Ⅲ. 부록

1. 본 윤리기준에서 인공지능의 지위

o 본 윤리기준에서 지향점으로 제시한 '인간성을 위한 인공지능(AI for Humanity)'은 인공지능이 인간을 위한 수단임을 명시적으로 표현하지만, 인간 종 중심주의(human species-centrism) 또는 인간 이기주의를 표방하지는 않는다.

o 본 윤리기준에서 인공지능은 지각력이 있고 스스로를 인식하며 실제로 사고하고 행동할 수 있는 수준의 인공지능(이른바 강인공지능)을 전제하지 않으며 하나의 독립된 인격으로서의 인공지능을 의미하지도 않는다.

2. 적용 범위와 대상

o 본 윤리기준은 인공지능 기술의 개발부터 활용에 이르는 전 단계에 참여하는 모든 사회구성원을 대상으로 하며, 이는 정부·공공기관, 기업, 이용자 등을 포함한다.

3. 인공지능 윤리기준의 실현방안

o '인공지능 윤리기준'을 기본 플랫폼으로 하여 다양한 이해관계자 참여하에 인공지능 윤리 쟁점을 논의하고, 지속적 토론과 숙의 과정을 거쳐 주체별 체크리스트 개발 등 인공지능 윤리의 실천 방안을 마련한다.

참고 2 국내외 주요 AI 윤리원칙(25개) 주요내용

	제목	주체	수립목적	주요 원칙	주요 특징
1	Preparing for the Future of Artificial Intelligence ('16)	President National Science and Technology Council Committee on Technology (정부기관)	**美 정부의 입장**에서 AI 기술과 관련하여 나아갈 방향 제시한 **정부 보고서**	공공선, 공정성, 안전, 투명성, 이해가능성, **선을 위한 AI(AI for good), 인간 가치(Human values)**	▪ AI를 주요 성장동력으로 보고 **美 정부의 역할 강조** ▪ **윤리원칙 제시보다는** 국제인도법에 근거한 AI 무기체계 개발 등 **다양한 AI 관련 이슈를 제시하는데 초점**
2	Tenets ('16)	Partnership on AI (민간 연구소)	학계, 재계, 정책입안자 등 **다양한 주체들의 협력 도모**	**AI 혜택 최대화, 다양한 주체들 간 협력,** 사생활보호, 견고함, 해악금지, 설명가능성	▪ 학계 기업 정책입안자 등 **다양한 주체들 간 협력을 강조하고,** 이를 통해 **대중 교육 등** 추진할 것을 제안 ▪ 기술 혜택 최대화의 전제로 사생활 보호, 연구공동체의 책임, 견고성, 해악금지 등 제시
3	AI Policy Principles ('17)	Information Technology Industry Council (민간 협회)	**개발자에 대한 정부 차원의 지원** 및 공적 영역과 사적 영역의 협업 강조	안전과 제어가능성, 해석가능성, 인간 존엄성, **데이터의 대표성, 유연한 규제접근,** 기회의 평등	▪ 개발자의 입장을 강조, **정부의 규제나 개발자에 대한 정보공개 요구에 부정적** ▪ **다만 개발자에게도** 안전한 설계, 데이터 대표성 등 **높은 수준의 책임성 요구**
4	DeepMind Ethics & Society Principles ('17)	DeepMind (기업)	**사내에서 AI 연구 수행시 윤리적 고려사항** 제시	사생활 침해 금지, 평등, 도덕성, 포용성, 안전과 책무성, **거버넌스·규제**	▪ **연구자에게 필요한 윤리원칙과 체크**리스트를 제시하면서도 안전과 책무성을 **보장하는 거버넌스·규제 필요성 제기**
5	Asilomar AI Principles ('17)	Future of Life Institute (민간 연구소)	미국 보스톤의 비영리 연구단체인 삶의 미래 연구소(Future of Life Institute) 주관으로 작성한 윤리원칙	인권보장, 개인정보보호, 해악금지, 공공성, 데이터 관리, 책임성, 통제성, 투명성, **무기경쟁 회피**	▪ 스티븐 호킹·일론 머스크 등 **다수의 AI학자, 미래학자 및 산학연 관계자들이 서명** ▪ AI 기술 연구자, 정책 입안자, 관련 산업 종사자에게 필요한 **윤리원칙 제시**
6	AI at Google: Our Principles ('18)	Google (기업)	**구글 AI 개발자**에게 필요한 윤리원칙 제시	사회적 혜택 증진, **불공정한 편견 지양, 설명가능, 사생활침해 방지**	▪ **개발을 제한해야되는 AI 어플리케이션**으로 해를 끼치는 기술, 인명을 해하는 무기관련 기술, 국제 규약 위반 감시기술 등 **제시**
7	Microsoft AI principles ('18)	Microsoft's AETHER(AI and Ethics in Engineering and Research) (기업)	**MS AI 개발자**에게 필요한 윤리원칙 제시	공정성, 신뢰성 및 안전, 사생활 및 보안, 포용성, 투명성, 책무성	▪ **MS 사내 윤리강령 성격**이 강하며, 책임질 수 있는 AI와 이를 위한 교육 강조
8	OpenAI Charter ('18)	OpenAI (민간 연구소)	**AI 기술 연구자**에게 필요한 윤리적 태도와 원칙 제시	공공선, 해악금지, 안전 담보, AI개발 선두주자, 타 연구단체 협력,	▪ **연구자의 자유로운 연구 증진에 초점,** 고도로 자율적인 AGI(artificial general intelligence) 상정
9	Principles for Trust and Transparency ('18)	IBM (기업)	**IBM 직원들을 대상**으로 AI 연구를 위해 제시된 윤리원칙	**인간 지능 증강(augment),** 데이터 소유권, 국경간 데이터 이동, 투명성	▪ **AI는 인간을 대체하는 것이 아니라 증강(augment)하기 위한 것임을 명시** ▪ AI 사용 여부·시기, 학습 데이터 출처 고지 등 규정
10	The Montreal Declaration for a Responsible Development of AI ('18)	University of Montreal (민간 대학)	몬트리올 대학에서 개발된 사회적으로 책임 있는 AI 연구를 위한 윤리원칙	복지(well-being), 자율성 존중, 사생활 보호와 친밀성, 연대성, 민주적 참여, 공평, 다양성 포용, 사려	▪ **친밀성(intimacy), 사려(prudence), 지속가능한 발전 등 다른 가이드라인에 잘 등장하지 않는 원칙 제시** ▪ 윤리원칙 제시와 함께 서명으로 선언에 동참하도록 장려
11	지능정보사회 윤리가이드라인 ('18)	정보문화포럼 (정부기관)	인간 중심의 지능정보사회 구현	**이용자 주도성, 이용자/시민참여,** 공익, 공정성, 위험예방, 프라이버시 보호	▪ 지능정보기술 관련 개발자 및 공급자의 윤리의식 고취 및 이용자의 오남용 방지 지침 ▪ **주체별(개발자, 공급자, 이용자) 세부지침 마련**
12	AI in the UK: Ready, Willing and Able? ('18)	영국 정부 (정부기관)	**영국 정부 차원에서 정책적으로 접근**할 수 있는 제언 제시	데이터 접근과 제어, 이해가능한 AI, **디지털 이해력 증진, 공중보건 관리**	▪ **영국이라는 특정 국가 입장**에서 공중보건 데이터 관리, AI 디지털 이해력 제고 등 **구체적으로 취할 수 있는 AI 관련 정책을 제시**

13	카카오 알고리즘 윤리헌장 ('18)	카카오 (기업)	**카카오 내 AI 관련 연구**시 지향되어야 할 윤리원칙 제언	사회윤리 준수, 차별 경계, 학습데이터 운영, 알고리즘 독립성 및 설명, 기술 포용성, **아동·청소년 보호**	▪ **국내기업 최초 AI 윤리헌장**으로, 알고리즘과 데이터에 대한 관리, **아동과 청소년에 대한 보호 필요성 등 강조**
14	The Future Computed: AI and its role in Society ('18)	Microsoft's AETHER (기업)	AI가 가져올 미래의 변화에 대응하기 위해 MS의 Aether 연구소에서 책자 제작	**AI에 의한 진보**, 공정성, 신뢰성 및 안전, 사생활 보안, 포용성, 투명성	▪ **AI가 경제사회적 진보를 이끌고 지역적 전지구적 문제를 해결할 것이라는 관점** ▪ AI가 직업과 직장에 미치는 영향에 공공부문과 민간부문이 협력해 대응할 필요성 제시
15	Discriminating Systems - Gender, Race, and Power in AI ('19)	AI Now (민간 연구소)	작업환경에서 다양성을 확보하기 위해 고려할 사항 제시	다양성, 해악금지, 개방성, 투명성	▪ 급여지급 기준의 인종별, 성별 공개, 직원 채용시 투명성 준수 등 제시 ▪ 특히 **AI 시스템 사용시 투명성·편견·해악에 대한 철저한 점검·감시·추적·공개를 강조**
16	Ethically Aligned Design(Ver. 2) ('19)	The IEEE Global Initiative on Ethics of Autonomous and Intelligent System (민간 학회)	IEEE에서 Ethics in Action 캠페인과 함께 아울러 공개된 보고서	인권, 복지우선, 책무성, 투명성, 오용의 인식	▪ 각 원칙별로 이론적 배경, 참고자료를 제시하고 윤리원칙뿐만 아니라 관련 분야들에 대한 자료 수록
17	이용자 중심의 지능정보사회를 위한 원칙 ('19)	방통위-KISDI (정부기관)	안전한 지능정보서비스 환경조성 및 이용자의 권리와 자유에 근거한 윤리원칙 제시	사람중심 서비스, 투명성과 설명가능성, 책임성, 안전성, 차별금지, 참여, **프라이버시와 데이터거버넌스**	▪ 안전한 지능정보서비스 환경조성 및 이용자 보호를 위해 모든 주체 사이의 협력 강조 ▪ 기업과 연구자들의 의견을 폭넓게 수렴하여 작성 ▪ **이용자 보호의 관점 강조**
18	로봇 윤리 기본 원칙(수정) ('19)	산업통상자원부 (정부기관)	**2007년에 만들어진 로봇윤리헌장을 수정 보완**	인간의 존엄성 보호, 공공선, 행복추구, 투명성, 제어가능성, 책무성, 안전성, 정보보호	▪ **로봇산업계에 종사하는 연구원, 개발자, 및 사용자**가 로봇과 AI을 설계·제작·공급·사용·관리하는 데 기준으로 삼는 가이드라인 제시
19	인간중심의 AI 사회 원칙 ('19)	일본 총무성 (정부기관)	**25명의 산학연 전문가로 구성된 '인간 중심의 AI 사회 원칙 위원회'**를 통해 제안	인간중심, 교육교양, 개인정보 보호, 보안, 공정경쟁, 공정성, 책임성, 투명성, 혁신	▪ **저출산, 고령화, 지방소멸, 재해 재난 등 일본이 처한 어려움을 AI가 해결**할 수 있을 것으로 상정 ▪ AI를 공공재로 활용하여 사회의 근본적인 변화와 혁신을 달성하여 지속 가능한 발전 추구
20	Ethics Guidelines for Trustworthy AI ('19)	EU (국가 정부기관)	**EU 산하의 50여명으로 구성된 AI 전문가 그룹** 주도	**인간 권리·자율성 보장, 기술적 견실성, 사생활, 데이터 관리,** 투명성, 다양성, 차별금지, 복지, 책무성	▪ **범국가 차원의 협업을 통해 신뢰할 수 있는 AI를 위한 윤리원칙 정립**에 초점을 맞춤 ▪ 각 원칙의 평가 리스트를 구체적으로 제시
21	Recommendation of the Council on AI ('19)	OECD (범국가 정부기관)	**OECD 디지털 경제 정책 위원회** 주관하에 제작	**포용적 성장, 지속가능 발전, 인간중심 가치,** 공정성, 투명성, 설명가능성, 견고성, 보안 및 안전, 책무성	▪ 윤리원칙 뿐 아니라 정책 입안자들에 대한 제언 제시, **국가별 정책 수립과 국제적 협력 도모**
22	The global landscape of AI ethics guidelines ('19)	Jobin. A., Ienca, M. & Vayena, E. (개인)	전 세계의 주요 84개의 AI 윤리 가이드라인을 분석	투명성, 정의, 해악금지, 책임, 사생활보호, 혜택 추구, 자유, 신뢰, 지속가능성, 연대성	▪ **주요 윤리원칙을 빈도수별로 분석하고 주로 선진국을 중심으로 발표되고 있음을 밝힘**
23	Understanding artificial intelligence ethics and safety ('19)	The Alan Turing Institute (국영 연구소)	**영국의 국영 연구소**인 Alan Turing 연구소에서 제작	존중, 연결, 보호, 돌봄, 공정성, 책임성, 지속가능성, 투명성	▪ AI 기술이 **데이터를 처리할 때 발생할 수 있는 위험이나 문제점을 예방하는 데 필요한 윤리원칙에 초점**
24	Principles Artificial Intelligence: A Map of Ethical and Rights-Based Approaches ('20)	Berkman Klein Center For Internet & Society (민간 연구소)	36개의 윤리 가이드라인에 등장한 윤리원칙들을 주제별로 분석	사생활보호, 책무성, 안전과 보안, 투명성과 설명가능성, 공정성과 차별금지, 인간의 기술통제, 전문적 책임	▪ **정부, 정부 기관, 사적 기관 등 다양한 주체들이 제시한 윤리원칙들을 8개의 주제로 분류하고 분석**
25	Rome Call for AI Ethics ('20)	로마 교황청 (민간기관)	**로마 교황청**에서 인간의 혁신적인 미래를 위한 AI 윤리원칙 제정	투명성, 포용성, 책임성, 불편부당성, 신뢰성, 보안과 사생활 보호	▪ **종교 기관인 가톨릭교회에서 제정**한 윤리 원칙으로, **인간 가족(human family)에 대한 봉사, 젊은 세대에 대한 준비, 자연의 회복 필요성** 등 제시

참고 3 「인공지능 윤리기준」(안) 의견수렴 결과 (요약본)

	주요 의견	검토
구조	**① '4단 3강 15륜' 유교적 명칭에 대한 우려** • 유교 원리의 원류가 한국이 아니며, 해당 개념은 중국을 떠올리게 할 수 있음 • 글로벌 차원에서 해당 프레임을 통해 인공지능 원칙을 설명하는 것에 언어적 한계 • 삼강오륜에 관한 내용이 세대에 따라서는 일반적인 개념은 아니므로, 오히려 이용자에게 선입견을 주거나 어려운 개념으로 오해하여 규범 수용을 어렵게 만들 수 있음 ※ 전문가 9명 의견	☞ 한국의 고유성 확보를 위해 해당 프레임을 도입하였으나, 중국 성리학을 연상하게 할 가능성이 있고, 경우에 따라 역효과를 가져올 수 있다는 점을 고려하여 **유교적 명칭을 제외**
	② 4단 삭제 및 구조변경 필요 • 4대 속성(4단)과 3대 원칙(3강) 내용의 경계가 모호하여 혼란을 초래할 수 있으므로 4단 삭제 필요 • 각 단계의 차이가 명확히 드러나지 않으며, 상하 관계가 분명하지 않음 • 구조가 복잡하고 직관적이지 않아 의미의 전달이 어려울 수 있으므로, 정확한 메시지 전달 차원에서 구조변경 고려 필요 ※ 전문가 16명 의견	☞ 'AI for Humanity'를 위해 갖추어야 할 4대 속성을 '구조'에 포함하여 제시하였으나, 해당 내용이 3대 원칙과 유사한 측면이 있음을 고려하여 혼란을 줄이기 위해 **전체 구조에서 4단 삭제** - 다만, 구조에서는 삭제하더라도 본 윤리기준의 목적인 '인간성을 위한 AI' 구현에 대한 **이해를 돕기 위해** 4단 속성에 포함된 내용을 윤리기준 **전문 등에 포함**
	③ 위계적 구조 및 범주화에 대한 우려 • 4단 3강 15륜 구조가 형식적으로는 깔끔해 보이지만 실질적으로 정당화하기 어려울 것으로 판단. 여러 가치가 어떤 위계를 갖는지에 대해 논쟁이 많을 것으로 예상	☞ 3강 15륜에 해당하는 용어가 모두 '원칙'으로 병기되어 혼란을 초래하므로, 명확히 구분되도록 **명칭 변경** ☞ 또한 원칙 재검토를 통해 **구조상 범주화를 제외**하고 계층적이지 않은 구조로 **재구성**

	• 3대 원칙 상호 간, 15개 실행원칙 상호 간 관련성이 일부 존재할 수 있다는 점에서 15개 실행원칙을 3대 기본원칙의 범주에 따라 분류하는 것은 적절치 않음. 따라서 원칙 상호 간 범주화를 하지 않는 방안을 건의 ※ 전문가 9명 의견	- 3강→3대 **기본원칙**, 15륜→10대 **핵심요건**
기본원칙	**① 기본원칙, 실행원칙 간 개념이 중복** • 원칙 간 상하 종속관계, 선후관계 등을 고려하여 원칙에 대한 정의를 내리고, 3대 기본원칙과 15개 실행원칙의 관계를 명확히 해야 효과적임 • 기본원칙과 하부 실행원칙의 개념이 중복적인 것 같음 ※ 전문가 11명 의견	☞ 3강 15륜에 해당하는 용어가 모두 '원칙'으로 병기되어 혼란을 초래하므로, 명확히 구분되도록 **명칭 변경을 변경하고 체계 재구조화** - 3강→3대 **기본원칙**, 15륜→10대 **핵심요건**
	② 실제 현장에서 적용하기 힘든 내용 존재 • "어떠한 경우라도"와 같은 강하고 절대적 기준이 적용될 경우, 현실적으로 AI 기술의 개발과 상용화가 어려울 수 있음 • 또한, AI 로봇 기술의 개발과 상용화를 금지한다는 뜻으로 이해될 가능성이 있음 • 영업비밀 이슈 등으로 현실적으로 실현하는 데에 한계가 있음 ※ 전문가 7명 의견	☞ 현실성을 고려하여 각 원칙에 대한 설명 중 일부 강한 표현의 **수위를 조정하되, 기**업·시민단체·학계 간 상충된 의견 모두를 종합적으로 고려하여 **절충된 내용**을 포함하도록 노력
	③ 일부 추상적 표현의 구체화 필요 • "위해가 되지 않을..." 같은 표현은 큰 의미로서 이해가 가나, 실질적 가이드로서 역할을 하는데 현실적인 어려움이 있을 것으로 예상 • 일부 개념이 다소 모호한 느낌. 기본원칙	☞ **본 윤리기준의 목적이** 모든 사회구성원에 대해 AI 개발 및 활용 **전 단계에서 통용가능한 기본적·포괄적 기준을 제시**하는 것이기 때문에 **각 원칙이 추상적인 성격을 가질 수밖에 없음을 감안할 필요** ☞ 각 원칙의 가치, 활용 목적이 보다 분명히 드러

	이라는 점에서 좀 더 쉽게 이해할 수 있는 개념 제시가 효과적일 것 ※ 전문가 12명 의견	나는 방식으로 **서술 체계를 보완**하여 이해를 돕고, - **향후 개발자, 제공자, 이용자 등 주체별 체크리스트를 마련하여 실질적인 윤리기준 실천 가이드로서의 역할**을 할 수 있도록 지원 예정
핵심요건	**① 실행원칙이 너무 많으며, 원칙 간 관계 불분명** • 일부 실행원칙이 3대 기본원칙에 부적합하게 분류되어 있다고 생각 • 15개 실행원칙은 너무 많으며, 원칙 모두가 사회적으로 수용되기는 어려움. 각각의 내용은 적절하게 구성되었으나, 원칙 간 연결 관계에 대한 재검토 필요 • 원칙 간 상하관계가 분명하지 않고 중첩되는 부분이 존재. 그리고 원칙 간 충돌 또는 모순이 발생할 경우 어떻게 해결할지 제시하지 않고 있음 ※ 전문가 12명 의견	☞ 3강 15륜 **체계를 재구조화**하여 실행원칙을 기본원칙 범주에 따라 분류하지 않도록 보완 ☞ 3강 15륜에 해당하는 용어가 모두 '원칙'으로 병기되어 혼란을 초래하고, 과도하게 느껴질 수 있으므로 **기본원칙과 핵심요건으로 명칭 구분** ☞ 기본원칙을 실천하고 이행할 수 있도록 AI 전체 생명 주기에 걸쳐 충족되어야 하는 본연의 의미를 살려 **핵심요건**이라는 표현 사용 ☞ 원칙 간 상충, 갈등 또는 윤리적 딜레마 발생 가능성과 해당 문제를 민주적 과정을 통해 해결해야 함을 **전문에 명시** (ex. EU의 신뢰할수 있는 윤리 가이드라인에서도 동일한 방식으로 해당 문제 해결 노력)
	② 실행원칙 상호배타성 점검 필요 • 15개의 많은 실행원칙을 나열하고 있어 중복적으로 보임 • 일부를 통합하는 방향을 고려하여 원칙을 좀 더 분명하게 하면 좋을 것 ※ 전문가 12명 의견	☞ 기존 원칙 간 중복성을 재검토하여 핵심요건을 간명화(15대 원칙 → 10대 요건) - 기존 '다양성 존중 원칙', '개방성 원칙', '포용성 원칙'을 **'다양성 존중' 요건으로 통합** - 기존 '통제성 원칙', '안전성 원칙'을 **'안전성' 요건으로 통합**
	③ 윤리기준 이행 주체가 불분명함 • "AI는..."으로 시작하는 것은 윤리기준을 이행할 주체에 관해 혼란을 야기할 수 있으며, AI가 독자적이고 독립적인 주체로 보여질 수 있음 ※ 전문가 5명 의견	☞ AI 개발부터 활용에 이르는 전과정에서 관련되는 **정부·공공기관, 기업, 이용자 등 모든 관련 주체를 의미**하나, 혼란을 최소화하기 위해 **각 원칙·요건에 대한 설명 문구 보완** - 특히 **AI가 독립적 주체로 오해되지 않도록 문장 주술구조 등 보완**

	• 일부 원칙에서 사용주체를 공급자, 사용자 등에 한정하지 말고 모든 이해관계자를 넣어야함, 이행주체에 공공기관과 정부도 포함 ※ 전문가 2명 의견	
	④ 일부 실행원칙 삭제 필요 • **(행복추구 원칙)** 행복이라는 개념이 너무 포괄적이고 모호하며, 행복추구권이 독자적인 기본권이 될 수 있는지 법적으로도 논란이 있음. 행보추구는 본질적인 욕구이므로 굳이 실행원칙으로 만들 필요가 있는지 의문 • **(견고성 원칙)** AI가 새로운 환경을 인식하고 불리한 조건에서도 작동할 것을 규정하고 있는데, 이는 현재 인공지능 개발 수준에서 충족시키기 어려운 수준으로 판단됨 - 인공지능의 견고성을 기술적으로 지향하고 있지만 완성하지 못한 상황으로 현실적인 한계가 많음 ※ 전문가 17명 의견	☞ **(행복추구 원칙)** 다른 시행원칙 보다 추상적이며 포괄적인 개념. **광범위하게 적용 가능한 상위 원칙의 수준과 유사**하여 타 원칙과 중복될 우려가 있을 수 있기에, **해당 원칙 삭제** ☞ **(견고성 원칙)** 현재 인공지능이 가지는 **불확실성, 예측불가능성, 학습하지 못한 환경에 대한 대응 어려움 등 한계를 고려**했을 때, **현시점에서 참조하기 어려운 기준이라는 다수 전문가의 지적 수용** - 실제 **국내외 주요 원칙 가운데에서도 일부만 해당 원칙을 포함**하고 있으며(비교 대상 25개 중 6개), OECD에서도 내용적으로는 '안전성'에 관련하여 서술하고 있음을 감안하여 **해당원칙 삭제**
기타	**◈ 기타 각계 전문가 제기 의견** • **국제사회 협력 측면의 내용**이 포함되었으면 좋겠음 & 최근 국제적 추세가 **AI 윤리에서의 국제공조를 강조**하는 것임 ☞ '연대성' 요건에 국제사회 협력 필요성에 대한 문구로 "윤리적 AI 개발 및 활용을 위한 국제사회의 협력을 위해 노력해야 함"을 추가 • 교육에 관한 기준, 구체적으로 **디지털 시대를 대비한 기술교육에 대한 가치가 포함**되면 좋겠음, **AI 관련 기술교육 및 AI 이용 관련 정보제공 등에 대한 측면**에서의 고민도 함께 되면 좋겠음 ☞ '공공성' 요건에 AI 교육 관련 문구로 "AI의 순기능을 극대화하고 역기능을 최소화하기 위해 다방면의 교육을 시행하여야 함"을 추가 ☞ 접근성, 이용 정보 제공 등의 내용은 "다양성 존중" 요건에 포함되어 있음 • **공공성 측면의 원칙이 과도하게 강조될 경우** 해외의 AI 기업에 비해 **국내 기업들의 경쟁력**이 떨어질 수도 있을 것을 우려, **현재의 윤리기준이 정책화(규제)될지 모른다는 우려** 존재. 과도한 불안감 조성이나 규제로 이어지는 것은 경계 필요	

☞ 본 윤리기준은 강제력·구속력 있는 '법' 또는 '지침'이 아닌, 기업의 자율성을 존중하고 기술발전을 장려하며 기술과 사회변화에 유연하게 대처할 수 있는 윤리 담론을 형성하기 위한 기준

☞ 또한, 본 윤리기준은 산업·경제 분야의 자율규제 환경을 조성하여 인공지능 연구개발 및 산업 성장을 저해하지 않고 정당한 이윤을 추구하는 기업에 부담을 지우지 않음을 목표로 한다는 내용을 윤리기준 서문 등에 명시

- AI 개발과 활용 전 단계에 걸쳐 참조할 수 있는 기준을 수립하고자 하면서도, 기업 자율성과 기술발전 장려를 너무 고려한 나머지 **아무런 규범력을 갖지 못하는 추상적인 '윤리기준'의 형태를 취한 것이 적절한지** 재고해 볼 필요가 있음, 이미 윤리기준은 많이 있으며, 윤리기준이 아닌 현실에서 윤리규범이 어떻게 이행되도록 할 것인지, 이를 위해 **어떠한 법규나 지침이 필요할 것인지 논의** 필요

☞ AI 기술의 특징 및 역기능 등에 대한 충분한 검증·평가가 이루어지지 않은 상황에서 법적 규제를 강제할 경우 정당한 AI 연구개발 및 산업성장이 저해될 수 있으며, 그 제정·시행 등이 경직적인 법 규범만으로는 급변하는 기술 환경 변화에 대응하는 데 한계도 있음. 그러므로 기술·사회 변화에 유연하게 대처할 수 있는 자율규제 환경조성을 위해 윤리기준 정립이 필요한 상황

☞ 동 윤리기준을 통해 인공지능 윤리 이슈에 대해 우리 사회 모두가 참조할 수 있는 기준을 제시하고 관련 논의가 발전해나갈 수 있는 플랫폼을 제공하며, 개별 분야에서 명백히 법제화가 필요한 쟁점들에 대해서는 사안별로 사회적 합의를 토대로 대응해 나갈 예정

- 엄격한 차원에서 일회적인 내용 제시에 그치는 것이 아니라, **다양한 방법과 채널을 통해 윤리기준이 지향하고 있는 본질적인 의미에 대해 이해할 수 있는 기회**가 개발자를 비롯한 일반 국민들에게도 반복적, 지속적으로 제공되었으면 함, 이번을 시작으로 지속적으로 논의가 이뤄지는 체계를 만들어 **AI 윤리를 지속적으로 논의** 할 수 있었으면 함, 관련 규범 형성 과정에 **기업들의 참여와 의견 수렴 과정을 지속 포함**하여 주시면 감사하겠음

☞ 본 윤리기준은 일회성으로 끝나는 것이 아닌, 이후 사회경제, 기술의 변화에 따라 새롭게 제기되는 인공지능 윤리 이슈에 대한 지속적 논의와 소통의 장인 플랫폼으로서의 기능을 할 것

☞ 공청회 개최 등을 통해 더 많은 개발자, 일반 국민들이 본 윤리기준에 대해 공감하고 본연의 의미를 이해할 수 있도록 노력할 예정

- **데이터 편향성, 알고리즘에 의한 차별 등 문제**는 **기술적으로도 해결이 가능**하므로 **관련 연구개발 투자 등이 필요**하다고 생각

☞ 윤리적 인공지능을 위해 신뢰할 수 있는 인공지능 속성 규격서 도출 등 관련 투자에 지속 노력 예정

- **인공지능 윤리기준 마련 과정**에서 **다양한 주체의 의견을 수렴해온 과정에 매우 모범적**이며 바람직하다고 생각. 앞**으로도 기업, 시민 단체의 폭넓은 참여가 있을 수 있도록 노력**해주시기 바람. 또한 인공지능 윤리기준이 **기술 고도화, 연구 결과 등 반영해 지속적으로 업데이트**되기를 바람.

☞ 윤리기준 발표가 1회성에 그치지 않고 새로운 이슈를 반영하여 지속 발전하고 학계, 기업, 시민단체 등 다양한 주체들이 인공지능 윤리이슈에 대해 논의하고 숙의하는 플랫폼이 될 수 있도록 운영해 나가겠음

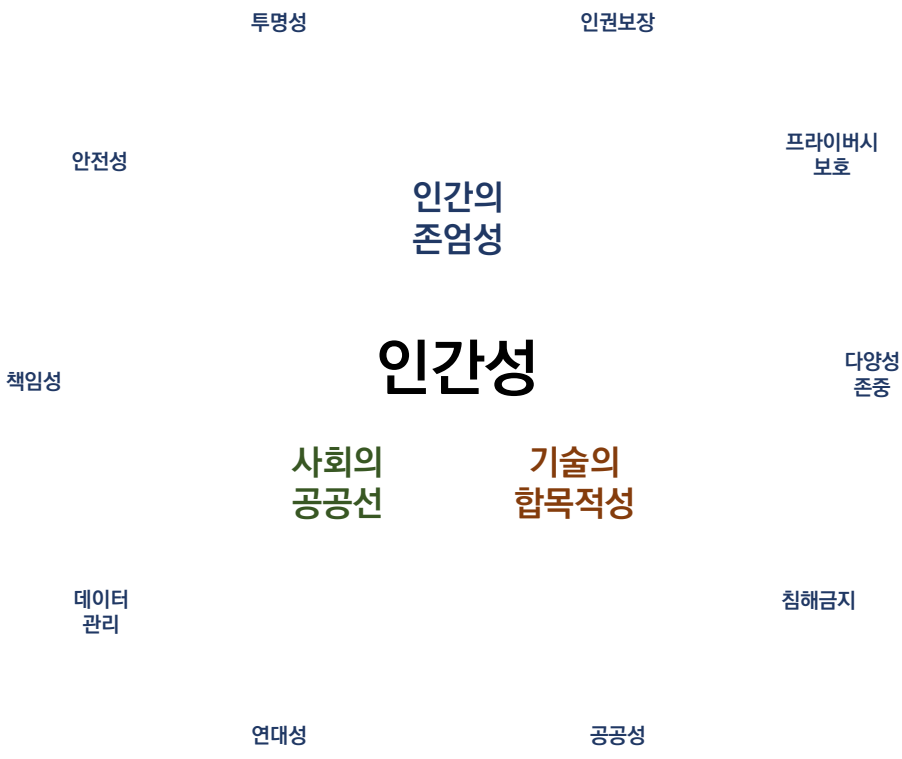
참고 4 인공지능 윤리기준 초안 vs. 최종본 비교

□ 인공지능 윤리기준 구조

초안

▼

최종본



□ 초안 · 최종본 내용 대비표

초안	최종본
	<공통 수정사항> ① 주체가 모호한 윤리기준 내용 내 **주어 명확화** ② 일부 불필요한 수동적 표현 → **능동적 표현** 수정 ③ 개조식 문장을 선언 형태의 **서술식 문장**으로 수정
AI가 최고 지향하는 가치 - 인간성(AI for Humanity)	**인공지능이 최고 지향하는 가치** - (초안과 같음)
4단(四端) : AI 기본 속성 내지 본질 ① 인권보장 ② 공공성 증진 ③ 인간 능력의 향상 ④ 기술윤리적 좋음	**< 삭 제 >** ☞ (삭제이유) **4대 속성(4단)과 3대 원칙(3강) 간 내용의 경계가 모호**하여 혼란을 초래할 수 있으므로 기준(안)에서 삭제하고 윤리기준 전문 등에 내용 반영
3강(三綱) : 기본원칙 ▪ AI가 규범적 의미의 인간성을 구현하기 위해 개발 및 활용과정에 고려해야할 기준 내지 원칙)	**3대 기본원칙** ▪ **인공지능 개발 및 활용 전 과정에서 고려할 기준** ☞ (수정이유) 향후 **국제적 논의를 고려**했을 때, 4단, 3강, 15륜 표현은 오해의 소지가 있으므로 3강→**3대 기본원칙**으로 수정 - 3대 기본원칙 상호 간, 15개 실행원칙 상호 간 관련성이 일부 존재할 수 있다는 점에서 **15개 실행원칙을 3대 기본원칙의 범주에 따라 분류하는 것은 적절치 않아 보인다**는 의견을 반영하여 재검토
3-① 인간의 존엄성 원칙 ▪ 신체와 이성을 가진 생명체로서의 인간은 AI를 포함하여 인간을 위해 개발된 기계 제품들과는 교환 불가능한 가치를 가짐 ▪ **보다 나은 삶을 위해 개발되고 만들어지는** AI는 인간의 생명은 물론 **정서적** 건강에 위해가 되지 않는 범위 내에서 개발되고 활용되어야 함 ▪ **AI는 개인의 사생활을 보호해야 하며, 어떠한 경우에라도 인간의 신체에 위해가 되지 않을 수 있는 안정성과 내구성을 가져야 함**	**3-① 인간의 존엄성 원칙** ▪ **인간은** 신체와 이성이 있는 생명체로 인공지능을 포함하여 인간을 위해 개발된 기계 제품과는 교환 불가능한 가치가 있다. ▪ 인공지능은 인간의 생명은 물론 **정신적 및 신체적** 건강에 해가 되지 않는 범위에서 개발 및 활용해야 한다. ☞ (수정이유) 불필요한 수식어구는 삭제하고 범위 명확화 ▪ **인공지능 개발 및 활용은 안전성과 견고성을 갖추어 인간에게 해가 되지 않도록 해야 한다.** ☞ (수정이유) "AI는…보호해야 하며"는 하위 요건에서도 서술하고 있어 삭제하며, 해외 주요 원칙에서 "안전성과 견고성"이라는 표현을 더 많이 활용(예: OECD, EU)하고 있음을 고려

3-② 사회의 공공선 원칙 ▪ **AI는 최대한 많은 사람의 안녕과 행복에 도움이 될 수 있도록 개발되어야 함** ▪ **인류의 삶을 위한 도구로서의** AI는 **최대한 많은 사람에게 최대한 많은 유익이 공평하게 분배될 수 있도록** 개발 및 활용되어야 함 ▪ 사회적, 국가적 나아가 **지구적 관점에서 공공의** 복지를 향상시킬 수 **있도록 AI는 개발 및 활용되어야 함**	**3-② 사회의 공공선 원칙** ▪ **공동체로서의 사회는 가능한 한 많은 사람의 안녕과 행복이라는 가치를 추구한다.** ☞ (**수정이유**) 원칙의 가치를 우선 설명하고, 후에 활용목적을 설명하는 방식으로 서술 체계 보완 ▪ 인공지능은 **지능정보사회에서 소외되기 쉬운 사회적 약자와 취약 계층의 접근성을 보장하도록** 개발 및 활용되어야 한다. ☞ (**수정이유**) “최대한~” 표현이 주는 모호함(예: 범위에 소수가 포함되지 않을 수 있음)을 해소하고, 본 원칙의 목적이 명확히 드러나도록 수정 ▪ **공익 증진을 위한 인공지능 개발 및 활용은** 사회적, 국가적, 나아가 **글로벌 관점에서 인류의 보편적** 복지를 향상시킬 수 있어야 한다.
3-③ AI의 목적성 원칙 ▪ **사회적 필요성이나 기술적 필요성에 의해 주어진 목적에 맞추어 개발된 AI는 당초 목적에 맞게 개발 및 활용되도록 사회적으로 보호되어야 함** ▪ 인간의 삶과 인류의 번영을 위한 **AI 산업은** 진흥되어야 함 ▪ AI 연구 및 산업은 ‘인간의 존엄성 원칙(1원칙)’, ‘사회의 공공선 원칙(2원칙)’ 내에서 지원·육성되어야 함	**3-③ 기술의 합목적성 원칙** ☞ (**수정이유**) 기존 원칙 명칭의 부족한 직관성을 보완하고, 지향하는 가치의 명확화를 위해 명칭 변경 ▪ **인공지능 기술은 인류의 삶에 필요한 도구라는 목적과 의도에 부합되게 개발 및 활용되어야 하며 그 과정도 윤리적이어야 한다.** ☞ (**수정이유**) 사회적 보호 등 표현이 의미전달이 어렵다는 지적과, 인공지능이 독립적 인격이 아닌 도구임을 명확히 해야 한다는 지적을 반영하면서 원칙의 핵심내용을 보다 잘 드러낼 수 있도록 문구 수정 ▪ 인류의 삶과 번영을 위한 **인공지능 개발 및 활용을 장려하여** 진흥해야 한다. **< 삭 제 >** ☞ (**삭제이유**) 1, 2원칙에서 제시하는 목표에 부합하는 차원에서 활용되어야 하는 것은 당연하고, 제3의 기본원칙이라는 점에서 별도의 내용을 포함하는 게 바람직하는 의견을 수용하여 기준(안)에서 삭제
15(十五倫) : 실행원칙 ▪ 3대 기본원칙을 실천하고 이행할 수 있도록 각 기본원칙에 포함되는 개념들을 재정의하고, 이에 맞게 세부 실행원칙을 범주화	**10대 핵심요건** ▪ 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 전체 생명 주기에 걸쳐 충족되어야 하는 10가지 핵심 요건을 제시

	☞ (**수정이유**) 기존 기본원칙, 실행원칙 용어가 모두 '원칙'으로 병기되어 혼란을 초래하므로 기본원칙과 핵심 요건으로 수정 ☞ (**수정이유**) 기존 원칙 간 중복성 재검토를 통해 원칙 및 핵심요건을 간명화 **①행복추구 → <삭제>** **③개인정보보호 → ② 프라이버시 보호** **④다양성 존중+⑦개방성+⑨포용성 → ③다양성 존중** **⑫통제성+⑬안전성 → ⑨안전성** **⑮견고성 → <삭제>**
15-① 행복추구 원칙 ▪ AI는 인간이 본질적으로 지닌 욕구이자 권리인 행복을 충족시키는 데에 도움이 되어야 함 ▪ AI는 인간의 행복추구 성향을 존중하고 삶의 질 향상을 통한 복지를 증진함으로써 인간 삶의 만족도 제고에 기여해야 함	**< 삭 제 >** ☞ (**수정이유**) 다른 핵심요건과는 달리 추상적이며 포괄적인 개념으로 기본원칙 수준으로 중복을 피하기 위해 삭제 ☞ (**수정이유**) 광범위하게 적용 가능하여 타 요건과 중복 우려
15-② 인권보장 원칙 ▪ 모든 인간에게 동등하게 부여된 **기본** 권리들을 존중하고, 다양한 민주적 가치들과 국제 인권법 등에 명시된 **기본권을** 보장하여야 함 ▪ **기본 권리들을 침해하는 방향으로 활용되어서는 안 되며, AI가 부정적으로 활용되거나 인간의 기본권과 자유를 침해할 것으로 예상되는 경우 이는 차단되어야 함**	**10-① 인권보장** ▪ **인공지능의 개발과 활용은** 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 **권리를** 보장하여야 한다. ☞ (**수정이유**) 인권 및 기본권 간 관계에 대해 발생할 수 있는 혼선을 최소화하고자 "기본권" 표현 수정 ▪ **인공지능의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.** ☞ (**수정이유**) "기본권" 표현 수정, 현재 인공지능 기술 수준을 고려해 현실적으로 가능한 수준에서 표현
15-③ 개인정보보호 원칙 ▪ **수집된 정보를 처리하는 과정에서 이용자 및 제3자의** 프라이버시를 침해하지 않도록 설계되어야 함	**10-② 프라이버시 보호** ☞ (**수정이유**) 개인정보보호는 이미 법률에서 규제하고 있는 사항으로, 더욱 큰 의미인 프라이버시 보호에 초점을 두기 위해 수정 ▪ **인공지능을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.** ☞ (**수정이유**) "처리하는", "이용자 및 제3자" 표현의 모호성, 한계성을 보완

▪ **기술적 차원에서는 개인정보 유출을 통한 사생활 침해를 방지하는 보안 장치가 마련되어야 함** ▪ **규범적 차원에서는** 정보보호에 관한 규범 등에 명시된 개인정보보호 및 정보 접근에 관한 규정을 준수하여야 함 ▪ **개인의 사생활과 관련된 민감한 정보(성적 지향, 정치관, 종교관 등)는 개인들을 편향되게 판단하거나 차별하는 근거로 활용되어서는 안 됨**	▪ < 삭 제 > ☞ (수정이유) 이미 관련법에서 해당 사항을 규제하고 있어 오히려 혼란을 줄수 있다는 전문가 지적 고려 ▪ < 삭 제 > ☞ (수정이유) 이미 관련법에서 해당 사항을 규제하고 있어 오히려 혼란을 줄수 있다는 전문가 지적 고려 ▪ **인공지능 전 생애주기에 걸쳐 개인 정보의 오용을 최소화하도록 노력해야 한다.** ☞ (수정이유) "개인의…활용되어서는 안 됨"과 관련된 내용은 "10-③ 다양성 존중" 요건에서 다루고 있어 중복이 없도록 삭제하고, 프라이버시 보호 측면에서 인공지능의 올바른 활용을 강조하도록 수정
15-④ 다양성 존중 원칙 ▪ AI는 정보 접근에 있어 다양성을 보장하는 한편, 다양성으로 인한 불공정한 대우를 최소화하도록 하는 데 사용되어야 함 ▪ **AI가 활용하는 데이터는** 사용자들의 다양성과 대표성이 반영되도록 **하고, 그에 따른 데이터의 무결성을 제공할 수 있어야 함** **< 15-⑦ 개방성 원칙 >** ▪ 성별·연령·장애·인종·종교·국가 등 개인특성에 따른 **편견이나 차별 없이 모든 사람이 AI를 공정하게 활용할 수 있어야 함** **< 15-⑨ 포용성 원칙 >** ▪ 기술 및 서비스에 대한 사회적 약자 및 취약 계층의 접근성을 보장함으로써 **AI의 혜택이 폭넓게 배분되도록 해야 함** ▪ AI로부터 얻는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 함	**10-③ 다양성 존중** ☞ (수정이유) 개인특성 등 다양성 존중, 접근성 보장, 차별 금지 등이 공통의 가치를 추구하는 측면을 고려하여 기존 다양성존중원칙, 개방성원칙, 포용성 원칙을 통합 ▪ < 삭 제 > ☞ (수정이유) 다양성 존중 요건 내 세부 요건과 중복되는 사항으로 삭제 ▪ **인공지능 개발 및 활용 전 단계에서** 사용자의 다양성과 대표성을 반영해야 하며, 성별·연령·장애·인종·종교·국가 등 개인 특성에 따른 **편향과 차별을 최소화**하고, **상용화된 인공지능은** 모든 사람에게 공정하게 **적용되어야 한다.** ☞ (수정이유) 타 핵심요건들과 볼륨을 맞추기 위해 공통적인 내용의 세부 요건 내용을 병합 ☞ (수정이유) 현재 인공지능 기술 수준을 고려해 현실적으로 가능한 수준에서 "무결성을 제공…" 표현 삭제 및 "최소화", "상용화된" 등의 표현으로 수정 ▪ 사회적 약자 및 취약 계층의 인공지능 기술 및 서비스에 대한 접근성을 보장하고, 인공지능이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다. ☞ (수정이유) 타 핵심요건들과 볼륨을 맞추기 위해 공통적인 내용의 세부 요건 내용을 병합

15-⑤ 해악금지 원칙	10-④ 침해금지
	☞ (수정이유) '해악' 표현이 형법상 협박죄에서 협박을 연상케 하는 점에서 좁은 의미로 생각할 수 있음. 따라서 보다 포괄적 의미의 침해금지로 수정
▪ **인간에 대해 미치는 해로움의 정도를 판단하고, 물리적·정신적 피해를 최소화하는 데 활용되어야 함**	▪ < 삭 제 >
	☞ (수정이유) 해당 내용의 전반적인 사항은 기본가치에 포함되어 있으므로 본 요건에서 해당 내용을 삭제
▪ AI는 인간에게 직·간접적인 해를 입히는 목적으로 활용되어서는 안 됨	▪ 인공지능을 인간에게 직간접적인 해를 입히는 목적으로 활용해서는 안 된다.
▪ AI가 야기할 수 있는 **위험의 정도 및 부정적 결과를 사전에 인지하고 평가할 수 있어야 함**	▪ 인공지능이 야기할 수 있는 **위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.**
	☞ (수정이유) 현재 인공지능 기술 수준을 고려해 현실적으로 구현 가능성을 고려하여 "사전에...있어야 함" 문장을 "대응 방안을...노력해야 한다"로 수정
▪ **AI의 편의성으로 인해 쉽게 발생 가능한 오용(misuse)을 최소화하기 위해 관련된 교육을 시행하여야 함**	▪ < 삭 제 >
	☞ '교육 필요성'은 10-⑤ 공공성으로 이동
15-⑥ 공공성 원칙	**10-⑤ 공공성**
	☞ (수정이유) 개인행복추구, 사회적 공공성 증진, 인류 공동 이익 증대를 모두 중요한 목적으로 설명
	☞ (수정이유) 원칙 간 상충, 갈등, 긴장관계 발생 가능성은 전문 등에 명시. 윤리기준에서 직접적인 해결책을 제시하는 것은 한계가 있음을 설명. 이후 윤리 논의 플랫폼에서 다양한 이해관계자가 참여하는 민주적 과정을 통해 숙의할 필요
▪ AI는 개인적 **차원의** 행복추구 뿐만 아니라 사회적 공공성 **확보,** 인류의 공동 이익을 **극대화하는 방향으로** 활용되어야 함	▪ 인공지능은 개인적 행복 추구뿐만 아니라 사회적 공공성 **증진과** 인류의 공동 이익을 위해 활용해야 한다.
▪ AI는 긍정적 사회변화를 **증진시키는** 방향으로 활용되어야 함	☞ "극대화" 표현은 오해의 소지가 있으므로 필요 이상의 과도한 표현을 삭제, 간결화
▪ **< 신 설 >**	▪ 인공지능은 긍정적 사회변화를 **이끄는** 방향으로 활용되어야 한다.
	▪ **인공지능의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.**

	☞ (수정 이유) AI윤리교육의 공공성을 고려하여 교육 관련 내용을 '공공성'의 항목에 포함
15-⑦ 개방성 원칙 ▪ **성별·연령·장애·인종·종교·국가 등 개인특성에 따른 편견이나 차별 없이 모든 사람이 AI를 공정하게 활용할 수 있어야 함**	**< 삭 제 >** ☞ '10-③ 다양성 존중'에 통합
15-⑧ 연대성 원칙 ▪ 다양한 집단 간 관계의 연대성을 **회복**-유지하고, 미래세대를 위한 충분한 배려를 **고려하여** 활용되어야 함 ▪ **AI는 검증, 토론, 의견교환 등의 민주적 과정을 통해 공적 영역과 사적 영역 간 협업을 강화하는 매개체 역할을 수행하며,** 다양한 주체들의 공정한 참여 기회를 보장하도록 하여야 함 ▪ **< 신 설 >**	**10-⑥ 연대성** ▪ 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능을 활용해야 한다. ▪ **인공지능 전 주기에 걸쳐** 다양한 주체들의 공정한 참여 기회를 보장하여야 한다. ☞ (**수정 이유**) 간결성을 높이고 의미를 명확하게 전달하고자 핵심적인 내용만 남기도록 수정 ▪ **윤리적 인공지능의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.** ☞ (**신설이유**) 전문가 의견을 반영, 국제사회 협력 필요성 추가
15-⑨ 포용성 원칙 ▪ **기술 및 서비스에 대한 사회적 약자 및 취약계층의 접근성을 보장함으로써 AI의 혜택이 폭넓게 배분되도록 해야 함** ▪ **AI로부터 얻는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 함**	**< 삭 제 >** ☞ '10-③ 다양성 존중'에 통합
15-⑩ 데이터 관리 원칙 ▪ 개인정보, **환경정보** 등 각각의 데이터를 그 목적에 부합하도록 활용하고, 목적 외 용도로 활용하지 않아야 함 ▪ 데이터의 수집과 활용의 전 과정에서 데이터 편향성을 최소화할 수 있도록 데이터 품질을 관리해야 함 ▪ **확보된 데이터가 외부의 공격으로부터 안전**	**10-⑦ 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. ☞ (**수정이유**) 사람중심의 인공지능 구현을 위해 데이터 품질 뿐만 아니라, 그로인해 발생할 수 있는 차별 및 정보유출 등의 잠재적 위험 또한 관리할 필요가 있으므로 내용에 추가 ▪ **< 삭 제 >**

하게 보호될 수 있도록 해야 함	☞ (**삭제이유**) 외부의 공격으로부터 보호해야한다는 표현은 지나치게 구체적인 부담이라는 전문가 지적 고려
15-⑪ 책임성 원칙 ▪ AI 개발 및 활용과정에서 책임주체를 설정하여 발생 가능한 피해를 최소화할 수 있도록 노력해야 함 ▪ AI 설계 및 개발자, 서비스 공급자, 사용자 간의 책임을 명확히 해야 함	**10-⑧ 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.
15-⑫ 통제성 원칙 ▪ **AI 활용과정에서 명백한 오류 또는 불법 행위가 발생할 경우 사용자가 그 작동을 즉각적으로 제어 또는 정지할 수 있는 기능을 가지고 있어야 함** ▪ **이상 징후 발생 시 사용자가 쉽게 인지할 수 있도록 하며 이를 반드시 서비스 공급자, 관리자, 사용자에게 알려야 함**	**< 삭 제 >** ☞ '10-⑨ 안전성'에 통합
15-⑬ 안전성 원칙 ▪ **AI의 연구, 설계, 개발, 배포, 사용 등** AI 활용 전 과정에 걸쳐 안전**과 보안**을 보장할 수 **있어야 함** ▪ **사용 연한 내 전반에 걸쳐 안전하게 작동하도록 설계되어야 하며, 제작자는 사용 연한이 만료된 제품의 관리에 대한 매뉴얼을 개발단계에서부터 마련하여야 함** **< 15-⑫ 통제성 원칙 >** ▪ AI 활용과정에서 명백한 오류 또는 **불법 행위가** 발생할 경우 사용자가 그 작동을 **즉각적으로** 제어 **또는 정지**할 수 있는 기능을 **가지고 있어야 함** ▪ **이상 징후 발생 시 사용자가 쉽게 인지할 수**	**10-⑨ 안전성** ▪ 인공지능 **개발 및** 활용 전 과정에 걸쳐 **잠재적 위험을 방지하고** 안전을 보장할 수 있도록 **노력해야 한다.** ☞ (**수정이유**) 전문가 의견을 반영, 실현 가능성을 고려하여 세부 요건 내용의 수위를 조절 ▪ **< 삭 제 >** ☞ (**삭제이유**) 관리 매뉴얼 등의 내용은 지나치게 구체적임. 향후 체크리스트 개발시에 포함되어야 할 필요 ▪ 인공지능 활용 과정에서 명백한 오류 또는 **침해가** 발생할 때 사용자가 그 작동을 제어할 수 있는 기능을 **갖추도록 노력해야 한다.** ☞ (**수정이유**) "불법 행위" 표현으로 인해 본 여건의 내용이 법규제를 연상시킨다는 전문가 의견 반영 - 현재 인공지능 기술 수준을 고려해 현실적으로 구현 가능성을 고려하여 "즉각적으로", "정지할 수 있는 기능을 가지고 있어야 함" 문장의 표현수정 ▪ **< 삭 제 >**

있도록 하며 이를 반드시 서비스 공급자, 관리자, 사용자에게 알려야 함	☞ (**수정이유**) '즉각적', '반드시' 등 내용은 윤리 기준이 기보다 법규제를 연상시킨다는 지적 반영 ☞ (**수정이유**) 인공지능은 기술적으로 오류가 발생하지 않는 이상 스스로 정상 작동을 했다고 판단하므로 이상 징후를 인지하고 사용자에게 알리는 것이 현재 인공지능 기술로는 구현하기 어렵다는 지적 반영
15-⑭ 투명성 원칙 ▪ 사회적 신뢰 형성을 위해 AI의 투명성·설명 가능성을 높이기 위한 노력을 기울여야 함 ▪ 다만, 타 원칙과의 상충관계를 고려하여, 활용 상황에 적합한 수준의 투명성과 설명 가능성을 고려해야 함 ▪ AI 기반 제품 또는 서비스 제공시 AI가 활용되고 있음을 사전에 고지해야함 ▪ **법으로 규정된 이해관계자의 요청 시 AI의 입력값, 내부 프로세스, 동작의 종류 및 상태 등을 요청자가 이해할 수 있는 방식으로 표시 또는 설명할 수 있어야 함**	**10-⑩ 투명성** ▪ 사회적 신뢰 형성을 위해 인공지능의 투명성과 설명 가능성을 높이고, 타 원칙과의 상충관계를 고려하여 활용 상황에 적합한 수준의 투명성과 설명 가능성을 **높이려는 노력을 기울여야 한다.** ▪ (병합) ☞ 타 핵심요건들과 수준을 맞추기 위해 공통적인 내용의 세부 요건 내용을 병합 ▪ 인공지능기반 제품이나 서비스를 제공할 때 인공지능의 활용 내용과 **활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.** ☞ 사전고지 내용은 투명성 원칙에 포함될 필요(OECD 투명성 원칙에서 사전고지 및 상호작용 내용 포함됨) ▪ **< 삭 제 >** ☞ (**삭제 이유**) AI 입력값, 내부프로세스 공개 내용은 현실적으로 기업 수용이 어렵다는 지적 반영
15-⑮ 견고성 원칙 ▪ AI는 새로운 환경을 인식할 수 있어야 하며, 그 환경에서도 타당한 추론과 적절한 판단을 할 수 있어야 함 ▪ AI는 오용되거나 기타 불리한 조건에서도 적절하게 작동되어야 함	**< 삭 제 >** ☞ (**삭제이유**) AI 기술 수준을 고려했을 때, 현시점에서 참조하기 불가능한 내용이라는 다수 지적 수용 - 25개 윤리원칙 비교분석 대상 중 6개 기관만 견고성을 포함하고 있으며, OECD도 견고성 원칙을 포함하나 내용적으로는 '안전성'에 관련되어 있음을 고려