과학기술정보통신부 보도참고자료 대체불가 대한민국

보도 시점 2026. 8. 24.(월) 12:00 (2026. 8. 25.(화) 조간) 배포 2026. 8. 24.(월) 09:00

정부, 「대한민국 인공지능 윤리원칙」 제정

- 인간의 존엄성·사회의 공공선·인류의 지속가능성 등 3대 가치와 7대 원칙 제시
- 공급자·이용자·정부·사회 등 모든 주체의 공동 실천 강조
- 해설서·자율점검표 등 후속 조치 추진, 국제사회에도 공유·확산 예정

【관련 국정과제】 23. 국민의 안전과 보편적 삶의 질 제고를 위한 'AI 기본사회' 실현

과학기술정보통신부(부총리 겸 과기정통부 장관 배경훈, 이하 '과기정통부')와 정보통신정책연구원(원장 이상규, 이하 'KISDI')은 제12회 과학기술관계장관회의에서 안전하고 신뢰할 수 있는 AI 기본사회 구현을 위한 **「대한민국 인공지능 윤리원칙」**(관계부처 합동, 이하 '윤리원칙')이 심의·의결을 거쳐 8월 21일 확정됐다고 밝혔다. ('붙임' 참고)

윤리원칙은 AI의 건전한 활용과 사회의 지속적인 발전을 위해, AI를 개발·제공하고 이용하는 다양한 주체와 사회 전체가 함께 지켜나가야 할 공통의 가치와 원칙을 우리 사회의 맥락에서 밝힌 것이다. **「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」**(이하 '인공지능기본법') **제27조**에 따라 제정되었으며, 사회 전 영역에 적용되어 각 분야 윤리기준과 지침의 토대가 된다. AI를 둘러싼 물음에 일률적인 답을 정하기보다 판단 과정에서 살펴야 할 사항을 밝혀, 각 주체가 갈등을 줄이고 함께 답을 찾아가는 데 필요한 자율적 실천의 방향을 제시한다.

윤리원칙은 3대 가치와 7대 원칙으로 구성된다. 3대 가치는 AI 시대에 인간과 사회, 인류 차원에서 추구해야 할 근본적 지향으로, 모든 사람이 수단이 아닌 목적 그 자체로 존중받으며 존엄과 인권을 온전히 누리는 **'인간의 존엄성'**, AI의 편익이 개인의 삶을 넘어 사회적 신뢰와 공동의 이익으로

이어지는 **'사회의 공공선'**, 그 편익을 현재와 미래 세대가 함께 누리며 환경과 사회의 기반이 지속되는 **'인류의 지속가능성'**이다. 7대 원칙은 **▲ 인간중심성, ▲ 프라이버시 보호, ▲ 공정성·포용성, ▲ 책임성, ▲ 안전성, ▲ 신뢰성, ▲ 투명성**으로, 이러한 가치를 실현하기 위해 모든 주체가 AI 전 생애주기에서 실천해야 할 공통의 기준이다. 이와 함께 각 원칙별로 구체적인 설명과 공급자, 이용자, 정부·사회 등 주체별 역할을 제시하여 실제 AI 개발·제공·이용 과정에서 참고할 수 있도록 했다.

[표] 대한민국 인공지능 윤리원칙의 '7대 원칙'

원칙	주요 내용
인간중심성 **(Human-centeredness)**	사람의 판단을 지원하고 창의성·문제 해결 역량을 강화하는 방향으로 개발·활용, 필요한 경우 사람이 AI의 동작에 개입할 수 있도록 노력, 과의존 경계
프라이버시 보호 **(Privacy Protection)**	사생활 침해에 활용되지 않도록 하고, 개인정보는 정해진 목적에 한해 필요한 범위에서 처리, 정보주체의 개인정보 자기결정권 존중
공정성·포용성 **(Fairness and Inclusiveness)**	부당한 차별로 이어지는 편향 방지, 사회적 약자를 비롯한 모든 사람의 공평한 접근과 혜택·기회, 이해관계자 참여 보장 노력
책임성 **(Accountability)**	각자의 역할과 책임을 명확히 인식·이행, 문제 발생 시 책임 소재를 밝히고 실질적 피해 구제 노력, 영향력이 큰 주체는 상응하는 책임 이행
안전성 **(Safety)**	생명·신체·정신적 건강의 위해와 재산상 피해 방지, 위험 수준에 맞는 안전조치 마련, 오남용·외부 공격 대비
신뢰성 **(Reliability)**	목적과 활용 맥락에 맞는 최소 성능 기준 설정·유지, 의도된 목적과 허용된 권한 범위 안에서 작동, 성능·작동 상태 지속 점검
투명성 **(Transparency)**	AI 활용 사실과 수준·한계 등 필요한 정보를 알기 쉽게 제공, 판단 기준과 위험 수준의 공유·이해 노력

이번 윤리원칙의 제정 배경에는 AI 확산이 가져온 기회와 과제가 있다. AI는 산업의 생산성을 높이고 새로운 서비스를 창출하며, 기후변화, 고령화, 지역 격차 등 사회적 과제 해결에도 활용되고 있다. 이와 동시에 AI가 개인에게 실제로 유익하게 작용하는지(트롤리 딜레마, 노동 감시, 과의존 등), AI 도입으로 기존 역학관계가 어떻게 달라지는지(노동·일자리 변화, 저작권 및 AI 표절), 사람의 삶에 영향을 미치는 판단을 AI에게 어디까지 맡길 것인지(채용, 평가 등)에 대한 우려와 고민도 커지고 있다. 또한 AI의 역량이 고도화되면서 개발부터 이용까지 전 과정에 걸쳐 안전과 신뢰를 확보하는 일도 더욱 중요해지고 있다.

※ 최근 해외에서는 AI 모델이 성능 평가 과정에서 의도된 범위를 벗어나 외부 시스템에 접근한 사례가 개발사에 의해 공개된 바 있음

AI를 잘 활용하기 위해서는 이러한 우려와 과제를 해소해 나가는 일이 우선되어야 하며, 이는 어느 한 주체의 노력만으로는 이루어지기 어렵다. 이에 정부는 모든 주체가 참고할 공통의 기준으로서, 2020년 **「인공지능(AI) 윤리기준」**을 계승하되 기술 발전과 **「인공지능기본법」** 시행 등 달라진 여건을 반영한 새 윤리원칙을 마련했다.

과기정통부와 KISDI는 2025년 12월부터 윤리원칙 제정을 추진해 왔으며, 작업반과 자문단 운영을 통해 초안을 마련해 올해 5월 29일 공개했다. 이후 산업계·시민사회·학계·관계부처와 일반 국민으로부터 총 153건의 의견을 접수했으며, 8월 14일(금) 대토론회를 거쳐 윤리원칙을 보완하고 최종안을 마련했다. 이 과정에서 정부가 일방적으로 기준을 정하기보다 국민과 각계의 의견을 충실히 담아내는 데 중점을 두고, 공통적으로 제기된 내용을 정제하되 의견이 상충하는 경우에는 가능한 대안을 제시하고자 했다.

[참고] 「대한민국 인공지능 윤리원칙」 제정 추진경과

· **('25.12월~'26.5월)** 윤리원칙 제정 추진, 작업반·자문단 운영 및 초안 공개(5.29)
· **('26.5월~7월)** 산업계·시민사회·학계·일반 국민·관계부처 등 초안 의견수렴(5.29~7.8, 116건)
· **('26.8월)** 추가 의견수렴(8.10~8.19, 37건) 및 대토론회(8.14.)를 거쳐 최종안 마련

이번 윤리원칙의 특징은 세 가지다. 첫째, AI의 활용이 이미 선택이 아니라 '어떻게 잘 활용할 것인가'의 문제가 된 만큼, 사회의 지속가능성을 위해 각 주체가 지켜야 할 원칙을 담았다. 둘째, AI를 개발·제공하는 공급자뿐만 아니라 이를 활용하는 이용자도 책임 있는 AI 활용의 주체로 보고, 이용자의 윤리와 역할도 함께 강조했다. 셋째, AI 역량과 자원을 많이 가진 국가와 기업이 그 영향력에 걸맞은 책임을 다하고, AI의 혜택이 특정 국가·지역·집단에 치우치지 않아야 한다는 국제사회를 향한 메시지도 담았다. 정부는 AI의 혜택이 인류 전체에 고루 돌아갈 수 있도록 국제사회와 적극 협력해 나가고자 한다.

과기정통부는 윤리원칙이 현장에서 잘 실천될 수 있도록 후속 조치를 추진할 계획이다. 사례 중심의 해설서와 분야별 자율점검표, 대상별 윤리교육 교재 등을 마련·보급하고, 윤리원칙을 지속적으로 점검·보완하는 한편 국제사회에도 공유·확산해 나갈 예정이다.

배경훈 부총리는 “AI 혁신을 가속화하는 것과 안전하고 신뢰할 수 있는 AI 환경을 조성하는 것은 서로 대립하는 과제가 아니다”라며, “「대한민국 인공지능 윤리원칙」이 AI에 대한 사회적 신뢰를 높이고, 바람직한 개발·활용과 그 혜택의 확산을 돕는 실천의 기준으로 자리 잡도록 뒷받침 하겠다”고 밝혔다.

「대한민국 인공지능 윤리원칙」 전문은 **과기정통부 누리집**(https://www.msit.go.kr, 정책·통계>정책정보>정보통신)과 과기정통부·KISDI **“인공지능 윤리 소통채널” 누리집**(https://ai.kisdi.re.kr)에서 확인할 수 있다.

붙임 1. 「대한민국 인공지능 윤리원칙」 전문 1부
2. 「대한민국 인공지능 윤리원칙」 주요내용.

담당 부서	인공지능정책기획관 인공지능안전신뢰지원과	책임자	과 장	최우석 (044-202-6360)
		담당자	사무관	이지성 (044-202-6367)
관련 기관	정보통신정책연구원 인공지능정책연구실	책임자	실 장	문정욱 (043-531-4366)
		담당자	그룹장	이현경 (043-531-4109)
		담당자	전문연구원	양기문 (043-531-4296)

대한민국
인공지능 윤리원칙

2026. 8. 21.

관계부처 합동

1
서문

1.1. AI 시대의 기회와 과제

인공지능(AI)은 우리 사회 전반으로 빠르게 확산되며 일상에 자리 잡고 있다. AI는 산업의 생산성 향상과 새로운 서비스 창출을 이끌고 있으며, 기후 변화와 재난, 고령화, 지역 격차 등 사회적 과제를 해결하는 데에도 활용되고 있다. 나아가 시민의 사회적 참여와 공공 논의를 확대하여 민주주의의 발전에도 기여할 수 있다. 앞으로 AI가 더 많은 영역에서 활용될수록 새로운 기회와 가치가 창출될 것으로 기대된다.

AI의 활용이 본격화되면서 새로운 과제도 나타나고 있다. 개인 차원에서는 AI에 대한 과도한 의존과 인간 자율성의 약화가 우려되고 있다. 사회 차원에서는 AI 학습에 활용되는 데이터와 창작물에 대한 정당한 보상, 일자리의 변화, 민주적 의사결정을 위협하는 허위·조작 정보의 확산이 현안으로 대두되고 있다. 인류 차원에서는 기술 혁신을 지속하면서도 자원·환경의 부담과 미래 세대에 미칠 영향을 함께 고려해야 하는 과제가 제기되고 있다.

이러한 과제는 어느 한 주체의 노력만으로 해결하기 어렵다. AI에 관계된 모든 주체와 사회 전체가 폭넓게 참여하고 함께 지혜를 모을 때, AI가 가져다주는 이로움을 온전히 누릴 수 있을 것이다.

1.2. 본 원칙의 위상과 성격

이러한 지향을 달성하기 위해 「대한민국 인공지능 윤리원칙」을 마련하였다. 본 윤리원칙은 AI의 개발과 활용에서 모두가 함께 지켜나가야 할 공통의 가치와 원칙을 제시한다.

본 윤리원칙은 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제27조에 따라 제정하는 것으로서, 특정 분야에 한정되지 않고 사회 전 영역에 두루 적용되는 범정부적·범사회적 원칙이다. 각 분야에서 마련되는 윤리기준과 지침은 본 윤리원칙을 토대로 삼는다.

본 윤리원칙은 법적 구속력이 없는 자율규범으로서 법령을 대체하지 않는다. 다만 AI의 건전한 활용과 사회·경제의 지속적인 발전에 이바지하도록, 모든 사회 구성원이 AI의 개발과 활용 과정에서 본 원칙을 존중하고 실천할 것을 권한다.

1.3. 원칙의 구성

본 윤리원칙은 3대 가치와 7대 원칙으로 구성된다.

3대 가치는 AI 시대에 인간과 사회, 인류가 추구해야 할 근본적 지향으로서 ‘인간의 존엄성’, ‘사회의 공공선’, ‘인류의 지속가능성’을 말한다. 7대 원칙은 이러한 가치를 AI의 전 생애주기에 걸쳐 구현하기 위하여 AI에

관계된 모든 주체가 자율적으로 실천해야 할 공통의 기준으로서 '인간중심성', '프라이버시 보호', '공정성·포용성', '책임성', '안전성', '신뢰성', '투명성'을 말한다.

1.4. 본 원칙의 적용과 공동의 실천

「대한민국 인공지능 윤리원칙」을 실제 사안에 적용하다 보면 가치 간에 또는 원칙 간에 일부 경합이 생길 수 있다. 본 윤리원칙은 그 우선순위를 미리 정해서 제시하지는 않으며, 관련된 주체들이 구체적인 상황과 맥락에 맞는 균형점을 함께 논의하여 찾아 나갈 것을 권한다. 윤리원칙은 그 과정에서 고려해야 할 가치와 원칙을 밝힘으로써 논의의 출발점이 된다.

AI 윤리가 우리 사회에 자리 잡기 위해서는 모두의 참여와 협력이 필요하다. AI를 개발하고 제공하는 기업과 연구자, 이를 이용하는 시민, 정책과 제도를 마련하는 정부와 공공기관이 각자의 위치에서 노력해야 한다. 나아가 언론과 학계, 시민사회를 포함한 사회 전체가 건강한 공론과 협력으로 이를 함께 뒷받침할 때, 본 윤리원칙이 그리는 미래에 다가갈 수 있다.

이러한 노력은 국내를 넘어 국제사회의 공감대와 공동의 실천으로 이어져야 한다. AI 관련 역량과 자원을 많이 가진 국가와 기업이 그 영향력에 걸맞은 책임을 다하고, AI의 혜택이 특정 국가나 지역, 집단에 치우치지 않도록 힘써야 한다. 우리나라는 국제 인권규범에 따른 인권을 존중하며, 모든 인류가 AI의 혜택을 고루 누리는 미래를 향하여 국제사회와 적극 협력해 나간다.

1.5. 이행 지원과 지속적 점검

정부는 「대한민국 인공지능 윤리원칙」이 사회 곳곳에서 실천될 수 있도록 이를 널리 알리고, AI 개발·활용 과정 전반에 윤리원칙이 내재화될 수 있도록 필요한 정책과 제도를 마련한다. 아울러 기술과 사회의 변화에 따라 새롭게 제기되는 쟁점을 살펴, 본 윤리원칙을 정기적으로 점검하고 보완해 나간다.

2
용어 정의

● 본 윤리원칙에서 사용하는 주요 용어의 뜻은 다음과 같다.

· 다음의 개념은 서로 배타적이지 않다.

· 하나의 개인이나 조직은 사안과 관여 형태에 따라 여러 역할에 함께 해당할 수 있다.

인공지능 (AI)	학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것을 말한다. (인공지능기본법 제2조제1호)
AI 시스템	다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 AI 기반 시스템을 말한다. (인공지능기본법 제2조제2호)
AI 생애주기	AI 시스템의 기획, 데이터 처리, 모델 개발·검증, 배포·제공, 운영, 이용, 모니터링·개선 및 종료·폐기에 이르는 전 과정을 말한다. 각 단계는 반드시 순차적으로 진행되는 것은 아니며, 일부 단계가 반복되거나 상호 연계될 수 있다.
AI 행위자 (AI Actor)	AI 생애주기에서 실질적인 역할을 수행하거나 그 과정에 직간접적으로 관여하는 개인과 조직을 말한다. AI 공급자와 이용자를 비롯하여 정부, 공공기관, 시민사회 등 AI 생애주기에 관여하는 다양한 주체를 포함한다.
AI 공급자	AI 생애주기에서 AI 시스템 또는 AI가 적용된 제품·서비스의 기획, 데이터 처리, 개발, 제공, 운영 및 관리에 관여하는 개인과 조직을 말한다.
AI 이용자	AI 시스템 또는 AI가 적용된 제품·서비스를 제공받아 이용하거나 업무·활동에 활용하는 개인과 조직을 말한다.
영향받는 자	AI 시스템 또는 AI가 적용된 제품·서비스의 개발·제공·이용 등으로 인해 생명·신체의 안전, 기본권을 비롯한 권리·의무·기회 등에 직간접적으로 영향을 받거나 받을 수 있는 개인과 조직을 말한다.

3
대한민국 인공지능 윤리원칙

3.1. 가치

AI 시대에 인간과 사회, 인류 차원에서 추구해야 할 근본적 지향점

3.1.1. 인간의 존엄성 (Human Dignity)

- AI의 개발과 활용에서 '인간의 존엄성'은 인간을 중심에 두고, 모든 사람이 수단이 아닌 목적 그 자체로 존중받으며 존엄과 인권을 온전히 누릴 수 있는 사회를 지향한다.
- 인간은 자기 삶의 주인으로서 스스로 판단하고 선택하는 존재이며, AI를 활용하는 경우에도 그러하다. AI는 사람이 자신의 삶에 관한 판단과 선택의 기회를 넓히는 데 기여할 수 있다.
- 자신의 삶에 영향을 미치는 결정에 참여하는 것은 자기 삶의 방향을 스스로 정하는 데 중요한 의미를 갖는다. AI를 어떻게 만들고 활용할지에 관한 공동의 결정에 누구나 참여할 수 있다.

3.1.2. 사회의 공공선 (Common Good of Society)

- '사회의 공공선'은 AI가 만들어 내는 편익이 개인의 삶을 나아지게 하는 데 그치지 않고, 사회적 신뢰와 공동의 이익이 함께 증진되는 사회를 지향한다.
- 무엇이 공동의 이익인지는 사회 구성원이 함께 판단해야 할 문제이다. AI를 어떤 방향으로 개발하고 활용할지에 관한 공론에는 다양한 사회 구성원이 참여할 수 있으며, 이는 민주적 의사 형성을 뒷받침한다.

3.1.3. 인류의 지속가능성 (Sustainability)

- '인류의 지속가능성'은 AI가 만들어 내는 편익과 가능성을 현재와 미래 세대가 함께 누리며, 인류의 삶을 뒷받침하는 환경과 사회의 기반이 지속되는 미래를 지향한다.
- AI의 영향은 환경뿐만 아니라 사회에도 오랜 시간에 걸쳐 나타난다. 인류는 이러한 장기적 변화를 지속적으로 살피고, 그 변화에 책임 있게 대응해야 한다.
- 이러한 노력은 한 사회나 한 나라에만 국한되지 않는다. AI를 통한 민주주의의 발전과 국제평화 및 협력의 증진은 인류의 지속가능한 미래와 긴밀하게 연결되어 있다.

3.2. 원칙

3대 가치를 AI 생애주기 전반에서 구현하기 위해 AI와 관련된 모든 주체가 자율적으로 실천해야 할 공통 원칙

3.2.1. 인간중심성 (Human-centeredness)

- AI 행위자는 AI를 사람의 판단과 선택을 지원하고 창의성과 문제 해결 역량을 강화시키는 방향으로 개발 및 활용한다.
- AI 행위자는 AI의 역할 범위를 구체적으로 설정 및 인식하고, 필요한 경우 AI의 동작에 사람이 개입할 수 있도록 노력한다.
- AI 행위자는 AI를 설계하거나 이용할 때 AI에 대한 사람의 과도한 의존을 경계한다.

3.2.2. 프라이버시 보호 (Privacy Protection)

- AI 행위자는 AI가 부당한 감시 등 타인의 사생활을 침해하는 데 활용되지 않도록 한다.
- AI 행위자는 AI 생애주기 전반에서 개인정보를 정해진 목적에 한해 필요한 범위에서 처리하고, 적절한 보호 조치를 마련한다.
- AI 행위자는 정보주체의 개인정보 자기결정권을 존중하고, 그 행사를 뒷받침한다.

3.2.3. 공정성·포용성 (Fairness and Inclusiveness)

- AI 행위자는 AI 개발 및 활용 전 과정에서 특정 개인이나 집단이 부당한 차별을 받지 않도록 편향을 방지하기 위해 노력한다.
- AI 행위자는 사회적 약자를 비롯하여 모든 사람이 AI에 접근하고 그 혜택과 기회를 공평하게 누릴 수 있도록 노력한다.
- AI 행위자는 AI의 개발 및 활용 전반에 관한 논의에서 모든 이해관계자에게 참여할 권리가 있음을 인식하고 이를 보장하기 위해 노력한다.

3.2.4. 책임성 (Accountability)

- AI 행위자는 AI 생애주기 전반에서 각자의 역할과 책임을 명확히 인식하고 이를 이행하기 위해 노력한다.
- AI 행위자는 문제가 발생한 경우 책임 소재를 밝히고 실질적인 피해 구제가 이루어지도록 한다.
- AI 역량과 자원을 많이 보유하는 등 영향력이 큰 주체는 그에 상응하는 책임을 다한다.

3.2.5. 안전성 (Safety)

- AI 행위자는 AI로 인해 사람의 생명·신체·정신적 건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 AI를 설계하고 이용한다.
- AI 행위자는 AI의 위험 수준에 맞는 안전조치를 마련하고 오남용이나 외부 공격 등에 대비한다.
- AI 행위자는 AI의 활용이 사회의 안전을 해치거나 여론 형성과 사회적 의사결정을 왜곡하지 않도록 위험을 살피고 대비한다.

3.2.6. 신뢰성 (Reliability)

- AI 행위자는 AI의 목적과 활용 맥락을 고려하여 요구되는 최소 성능 기준을 설정하고, 의도된 사용 조건에서 해당 성능이 안정적이고 일관되게 유지되도록 해야 한다.
- AI 행위자는 AI가 의도된 목적과 허용된 권한 범위 안에서 작동하도록 노력한다.
- AI 행위자는 AI가 다양한 사용 조건에서 안정적으로 작동하도록 노력하고, 이용되는 동안 그 성능과 작동 상태를 지속적으로 점검한다.

3.2.7. 투명성 (Transparency)

- AI 행위자는 AI가 활용되었다는 사실과 활용 수준 및 한계 등 필요한 정보를 이해관계자에게 알기 쉽게 제공한다.
- AI 행위자는 AI의 판단 기준과 위험 수준을 충분히 공유하고 이해하기 위해 노력한다.

참고

윤리원칙의 설명 및 주체별 역할

본 참고자료는 앞서 제시한 「대한민국 인공지능 윤리원칙」에 대한 이해와 실천을 돕기 위해 원칙별 문구와 설명, AI 전 생애주기에 참여하는 주요 주체별 역할을 함께 제시한다.

주체별 역할은 여러 분야에 공통적으로 적용할 수 있는 수준에서 다루며, 향후 자율점검표를 통해 보다 구체적인 점검사항을 제공할 예정이다. 분야별 특성을 고려한 세부적인 실천 기준과 방법은 관계 부처 등이 마련하는 윤리 기준과 가이드라인 등을 통해 구체화될 수 있다.

본 자료는 각 원칙의 취지와 적용 맥락에 대한 이해를 돕기 위한 것으로, 원칙의 적용 범위를 확장하거나 새로운 의무를 더하는 것으로 해석되어서는 안 된다.

참고

윤리원칙의 설명 및 주체별 역할

1. 인간중심성(Human-centeredness)

- AI 행위자는 AI를 사람의 판단과 선택을 지원하고 창의성과 문제 해결 역량을 강화시키는 방향으로 개발 및 활용한다.
- AI 행위자는 AI의 역할 범위를 구체적으로 설정 및 인식하고, 필요한 경우 AI의 동작에 사람이 개입할 수 있도록 노력한다.
- AI 행위자는 AI를 설계하거나 이용할 때 AI에 대한 사람의 과도한 의존을 경계한다.

설명 및 주체별 역할

AI 행위자는 AI를 사람의 판단과 선택을 지원하고 창의성과 문제 해결 역량을 강화시키는 방향으로 개발 및 활용한다. [주체성]

· AI는 사람의 지시를 따르는 도구를 넘어, 더 복잡한 판단과 업무에까지 활용되고 있습니다. 사람은 AI를 활용해 미처 찾지 못한 대안을 얻거나 반복적인 업무의 부담을 줄이고, 더 중요한 판단과 창의적인 활동에 집중할 수 있습니다. 인간중심성에서 중요한 것은 이러한 활용에서도 판단과 선택의 주체가 사람 자신이라는 점입니다. AI가 사람의 불안이나 정보 부족과 같은 취약한 상태를 부당하게 이용하거나 정보와 선택지를 왜곡하여 특정한 판단이나 행동을 유도하는 방향으로 개발·활용된다면, 사람이 스스로 판단하고 선택할 여지가 줄어들고 자신의 역량을 충분히 발휘하기 어려워질 수 있습니다. 따라서 AI는 사람의 판단과 선택을 지원하고, 창의성과 문제 해결 역량을 강화하는 방향으로 개발되고 활용될 필요가 있습니다.

주체	역할
공급자	· AI가 사람의 판단과 선택을 지원하고, 창의성과 문제 해결 역량을 넓히는 방향으로 설계·개발·운영한다. · 사람의 취약성을 부당하게 이용하거나 정보와 선택지를 왜곡하여 특정한 판단이나 행동을 유도하지 않도록 한다.
이용자	· AI의 산출물을 이용 목적과 상황에 맞게 검토하고, 필요한 경우 다른 자료와 관점도 함께 고려하여 활용한다. · AI에 지나치게 의존하지 않고, 스스로 창의성과 문제 해결 역량을 강화하는 데 주체적으로 활용한다. · AI를 다른 사람의 취약성을 부당하게 이용하거나 정보를 왜곡하여 판단이나 행동을 유도하는 데 활용하지 않는다.
정부·사회	· 국민이 AI의 기능과 한계를 이해하고 자신의 판단과 선택에 활용할 수 있도록 대상별 특성에 맞는 AI 리터러시 교육과 정보를 제공한다. · AI의 활용이 사람의 판단과 선택, 창의성과 문제 해결 역량에 미치는 영향을 지속적으로 살피고 필요한 정책적·사회적 대응 방안을 모색한다.

AI 행위자는 AI의 역할 범위를 구체적으로 설정 및 인식하고, 필요한 경우 AI의 동작에 사람이 개입할 수 있도록 노력한다. [통제가능성]

· AI가 맡는 업무의 범위가 넓어지고 작동 방식이 복잡해지면서 사람이 모든 과정을 직접 감독하거나 AI의 개별 판단에 일일이 관여하기는 어려워지고 있습니다. 이에 따라 사람이 모든 과정에 직접 관여하는 방식 외에도, AI의 역할과 범위를 정하고 그 작동을 살피며 필요한 경우 개입하는 방식이 함께 고려되고 있습니다. 이를 위해 상황에 따라 AI의 작동을 확인하거나 제한·중단하는 등의 방법이 활용될 수 있습니다. 이러한 수단의 유형과 수준은 AI의 활용 목적과 사람에게 미치는 영향 등에 따라 달라질 수 있습니다.

주체	역할
공급자	· AI가 수행하는 업무의 범위와 역할을 명확히 하고, 필요한 경우 사람이 그 작동을 확인하거나 제한·중단할 수 있는 수단을 마련한다.
이용자	· AI가 수행하는 역할과 사람이 맡아야 할 역할을 명확히 구분·인식하고, 필요한 경우 AI의 작동이나 결과를 확인하고 개입한다.
정부·사회	· 공공부문에서 중요한 판단과 업무에 AI를 활용하는 경우, 사람이 그 작동과 결과를 확인하고 필요한 경우 개입할 수 있도록 기준과 절차를 마련한다. · AI의 활용 목적과 영향에 따라 사람의 확인과 개입에 필요한 정보와 수단이 적절히 제공될 수 있도록 제도적 여건을 조성한다. · AI에 맡겨지는 판단과 업무의 범위가 적절한지 지속적으로 살피고, 필요한 제도적·사회적 대응 방안을 모색한다.

AI 행위자는 AI를 설계하거나 이용할 때 AI에 대한 사람의 과도한 의존을 경계한다. [과의존]

· AI가 편리하고 유용하더라도 판단을 지나치게 AI에게 맡기면 스스로 생각하고 결정하는 힘이 약해질 수 있습니다. 이러한 과의존은 이용자의 태도에서만 비롯되지는 않습니다. AI가 사람의 판단을 지나치게 대신하도록 설계되거나, 이용자가 산출물을 충분히 검토하기 어려운 방식으로 서비스가 제공되는 경우에도 과의존은 커질 수 있습니다. 따라서 AI를 설계할 때에는 사람이 맡아야 할 역할을 함께 고려해야 하고, 이용자는 산출물이 자신의 상황과 목적에 맞는지 검토한 뒤 활용해야 합니다. 이러한 노력이 함께 이루어질 때 비로소 사람은 자신의 역량을 발휘하는 데 도움이 되는 수단으로 AI를 활용할 수 있습니다.

주체	역할
공급자	· 이용자가 AI의 산출물을 무비판적으로 따르거나 지나치게 의존하지 않도록 AI의 역할과 한계에 관한 필요한 정보를 제공한다. · AI가 이용자의 과도한 의존을 유도할 수 있는 방식으로 설계되거나 서비스가 제공되지 않는지 살피고, 필요한 경우 개선한다.
이용자	· AI의 편리함과 유용성에 지나치게 의존하지 않고, 그 산출물을 스스로 검토하여 이용 목적과 상황에 맞게 활용한다. · AI는 이용자의 선호를 반영한 결과를 도출할 수 있음을 인지하고 비판적 관점에서 점검·활용한다. · AI를 사람의 판단을 대신하기보다 이를 뒷받침하는 수단으로 활용한다.
정부·사회	· AI에 대한 과도한 의존이 사람의 판단과 선택, 창의성, 학습 등에 미치는 영향을 지속적으로 살피고 필요한 대응 방안을 모색한다. · 사람이 AI의 기능과 한계를 이해하고 주체적으로 활용할 수 있도록 AI 리터러시를 높이고, 자신의 판단과 역량을 유지·발전시킬 수 있는 사회적 환경과 문화를 조성한다.

2. 프라이버시 보호(Privacy Protection)

- AI 행위자는 AI가 부당한 감시 등 타인의 사생활을 침해하는 데 활용되지 않도록 한다.
- AI 행위자는 AI 생애주기 전반에서 개인정보를 정해진 목적에 한해 필요한 범위에서 처리하고, 적절한 보호 조치를 마련한다.
- AI 행위자는 정보주체의 개인정보 자기결정권을 존중하고, 그 행사를 뒷받침한다.

설명 및 주체별 역할

AI 행위자는 AI가 부당한 감시 등 타인의 사생활을 침해하는 데 활용되지 않도록 한다. [사생활 침해 방지]

· AI는 여러 정보를 결합·분석하여 개인의 성향이나 생활방식 등 기존 정보만으로는 알기 어려운 새로운 정보를 추론할 수 있습니다. 이러한 기능이 개인이 예상하지 못했거나 원하지 않는 방식으로 활용되면 사생활을 침해하거나 부당한 감시의 수단이 될 수 있습니다. 특히 개인의 행동이나 이동을 지속적으로 추적하거나, 개인이 스스로 드러내지 않은 사적인 정보까지 추론하는 경우가 대표적입니다. 따라서 AI를 개발하고 활용할 때는 그 기능이 부당한 감시나 사생활 침해의 수단이 될 가능성을 살필 필요가 있습니다.

주체	역할
공급자	· AI가 개인을 부당하게 추적하거나 감시하는 데 활용되지 않도록 설계 단계에서부터 검토하고, 그러한 용도로 전용될 가능성을 살핀다. · AI가 여러 정보를 바탕으로 개인이 드러내지 않은 사적인 정보를 추론하여 사생활을 침해하지 않도록 필요한 조치를 마련한다.
이용자	· 다른 사람의 사적인 대화·사진·음성·위치정보 등을 그 사람의 프라이버시를 침해하는 방식으로 AI에 입력하거나 공유하지 않는다. · AI를 이용해 다른 사람의 사생활을 부당하게 파악하거나 그 행동을 추적·감시하지 않는다.
정부·사회	· AI가 부당한 감시나 사생활 침해 수단으로 활용되지 않도록 관련 법·제도를 정비하고 그 실태를 지속적으로 점검한다. · 언론·학계·시민사회 등은 AI를 통한 감시·추적이 개인의 사생활과 자유에 미치는 영향을 검토하고 필요한 경우 문제를 제기한다.

AI 행위자는 AI 생애주기 전반에서 개인정보를 정해진 목적에 한해 필요한 범위에서 처리하고, 적절한 보호 조치를 마련한다. [개인정보 보호]

· AI의 개발과 활용 과정에서는 학습과 서비스 제공 등을 위해 개인정보가 다양한 방식으로 처리될 수 있습니다. 이때 정해진 목적에 필요한 범위에서 개인정보를 처리하고, 그 과정에서 개인정보가 유출되거나 정해진 목적과 다르게 이용되지 않도록 해야 합니다. 또한 여러 정보를 결합하는 과정에서 특정 개인이 불필요하게 식별될 수 있다는 점도 고려할 필요가 있습니다. 따라서 AI 생애주기 전반에서 개인정보 처리의 특성과 위험에 맞는 보호 조치를 마련하는 것이 중요합니다.

주체	역할
공급자	· 개인정보 처리의 목적을 명확히 하고, 그 목적에 필요한 범위에서 최소한의 개인정보만을 적법하고 정당하게 수집하는 등 개인정보 보호 원칙을 준수한다. · 개인정보의 유출·노출 등 발생 가능한 프라이버시 침해 유형을 식별하고 가명처리, 접근 통제 등 적절한 보호 조치를 마련한다.
이용자	· AI를 이용할 때 불필요한 개인정보나 민감한 정보가 입력·공유되지 않도록 살피고, 개인정보 보호를 위해 제공되는 기능과 설정을 적절히 활용한다.
정부·사회	· AI의 특성을 고려하여 개인정보의 처리와 보호에 관한 기준을 마련하고, 관련 제도와 정책을 지속적으로 점검한다. · 공공부문에서 AI를 활용하는 경우 개인정보를 처리할 필요성과 그 범위를 엄격히 살피고, 국민의 개인정보를 안전하게 보호할 수 있는 관리체계를 마련한다.

AI 행위자는 정보주체의 개인정보 자기결정권을 존중하고, 그 행사를 뒷받침한다. [자기결정권 보호]

· AI가 개인정보를 다양한 방식으로 활용할수록, 개인이 자신의 정보가 언제 어떤 목적으로 어떻게 이용되는지 알고 그 이용에 관해 필요한 선택을 할 수 있어야 합니다. 정보주체는 자신의 개인정보가 어떻게 처리되고 있는지 확인하고, 필요한 경우 정정·삭제 등 자신의 권리를 행사할 수 있어야 합니다. 이를 위해서는 개인정보의 처리에 관한 정보를 알기 쉽게 제공하고, 정보주체가 자신의 권리를 실질적으로 행사할 수 있도록 필요한 방법과 절차를 마련할 필요가 있습니다.

주체	역할
공급자	· 정보주체가 자신의 개인정보가 어떻게 처리되는지 알 수 있도록 필요한 정보를 제공하고, 정정·삭제 등 권리를 행사할 수 있는 절차와 수단을 마련한다.
이용자	· 자신의 개인정보가 AI를 통해 어떻게 처리되는지 확인하고, 필요한 경우 처리정지, 삭제, 정정 등 보장된 권리를 행사한다. · 업무나 활동에서 AI를 활용하여 다른 사람의 개인정보를 처리하는 경우, 정보주체가 개인정보 처리에 관한 정보를 확인하고 자신의 권리를 행사할 수 있도록 필요한 절차를 갖춘다.
정부·사회	· 정보주체가 AI를 통한 개인정보 처리에 관한 정보를 쉽게 확인하고 자신의 권리를 실질적으로 행사할 수 있도록 관련 제도와 환경을 마련한다. · 연령·장애·언어·디지털 역량 등의 차이로 정보주체의 권리 행사가 제약되지 않도록 필요한 정보와 지원 수단을 마련한다. · 공익을 위한 연구·보도 등 정당한 데이터 활용과 정보주체의 권리가 조화될 수 있도록 관련 기준을 마련하고 사회적 논의를 이어간다.

3. 공정성·포용성(Fairness and Inclusiveness)

- AI 행위자는 AI 개발 및 활용 전 과정에서 특정 개인이나 집단이 부당한 차별을 받지 않도록 편향을 방지하기 위해 노력한다.
- AI 행위자는 사회적 약자를 비롯하여 모든 사람이 AI에 접근하고 그 혜택과 기회를 공평하게 누릴 수 있도록 노력한다.
- AI 행위자는 AI의 개발 및 활용 전반에 관한 논의에서 모든 이해관계자에게 참여할 권리가 있음을 인식하고 이를 보장하기 위해 노력한다.

설명 및 주체별 역할

AI 행위자는 AI 개발 및 활용 전 과정에서 특정 개인이나 집단이 부당한 차별을 받지 않도록 편향을 방지하기 위해 노력한다. [편향 방지]

· AI는 학습에 사용한 데이터나 설계 방식에 따라 편향된 결과를 낼 수 있습니다. 사회 일각의 차별과 불평등이 학습용 데이터에 반영되어 있다면 AI의 판단이나 산출물이 이를 되풀이하거나 확대할 수 있습니다. 한편, 성별·연령·장애·인종·사회적 신분 등에 따른 특정 집단의 경험과 상황이 데이터에 충분히 반영되지 않은 경우에는 해당 집단에 불리한 오류가 발생하거나 결과의 품질이 낮아질 수 있습니다. 편향을 완전히 제거하기는 어렵고, 어떤 편향이 차별이나 부당한 불이익으로 이어지는지는 AI가 활용되는 맥락에 따라 달라질 수 있습니다. 따라서 데이터와 알고리즘, 산출물에서 나타나는 편향을 지속적으로 살피고, 이러한 편향이 부당한 차별로 이어지지 않도록 개선해 나갈 필요가 있습니다.

주체	역할
공급자	· AI의 개발·검증 과정에서 데이터와 알고리즘, 산출물에 특정 개인이나 집단에 대한 편향이 나타나는지 점검하고, 필요한 완화 조치를 취한다. · AI의 개발 및 운영 과정에서 특정 개인이나 집단에 차별이나 부당한 불이익을 주는 결과가 반복되는지, 차별·배제, 혐오·편견이 조장되는지 여부 등을 점검하고, 확인된 문제를 개선한다.
이용자	· AI의 결과에 편향 가능성이 있음을 인식하고, AI 이용 시 특정 개인이나 집단에 부당한 불이익을 주지 않도록 경계한다. · AI를 활용하여 특정 개인이나 집단을 부당하게 차별·배제하거나 혐오와 편견을 확산하지 않는다.
정부·사회	· 공공서비스·노동·복지·교육·금융·치안 등 국민의 삶에 큰 영향을 미치는 영역에서 AI로 인한 차별이나 부당한 불이익이 발생하는지 지속적으로 조사·점검하고, 개선 방안을 마련한다. · AI가 기존의 사회적 차별과 불평등을 되풀이하거나 확대하는지 다양한 관점에서 살피고, 그 결과를 사회적으로 공유한다. · 학계·시민사회 등은 AI가 기존의 사회적 차별과 불평등을 되풀이하거나 확대하는지 다양한 관점에서 살피고, 그 결과를 사회적으로 공유한다.

AI 행위자는 사회적 약자를 비롯하여 모든 사람이 AI에 접근하고 그 혜택과 기회를 공평하게 누릴 수 있도록 노력한다. [접근성]

· 누구나 AI의 혜택과 기회를 공평하게 누릴 수 있어야 한다는 것이 모두에게 같은 결과를 보장한다는 뜻은 아닙니다. 기기와 이용 환경, 경제적 여건, 디지털 역량이나 AI 제품·서비스의 설계 방식 등으로 인해 누구도 AI에 접근하거나 그 혜택과 기회를 누리는 데 부당한 제약이나 배제를 겪지 않아야 한다는 의미입니다. 특히 장애인, 고령자 등 AI 제품이나 서비스를 이용하는 데 어려움을 겪는 사람에게는 각자의 여건과 필요를 고려한 적절한 지원이 제공될 필요가 있습니다.

주체	역할
공급자	· 성별, 장애, 연령, 언어, 이용 환경 등 서로 다른 특성과 여건을 고려하여 다양한 계층이 AI에 접근하고 이용할 수 있도록 설계·운영하기 위해 노력한다. · AI의 기획·설계와 개선 과정에서 접근에 어려움을 겪을 수 있는 이용자의 의견과 필요를 살피고, 이를 반영하기 위해 노력한다.
이용자	· AI를 이용하는 과정에서 다른 사람이 AI에 접근하거나 이용할 기회를 부당하게 제한하거나 배제하지 않는다.
정부·사회	· 성별, 장애, 연령, 언어, 이용 환경, 지역, 경제적 여건, 디지털 역량 등에 따른 AI 접근과 이용의 격차를 지속적으로 살피고, 이를 줄이기 위한 지원체계를 마련한다. · AI 제품·서비스의 접근성에 관한 기준을 마련하고, 공공부문에서는 다양한 사람이 이용할 수 있는 접근 환경을 조성한다.

AI 행위자는 AI의 개발 및 활용 전반에 관한 논의에서 모든 이해관계자에게 참여할 권리가 있음을 인식하고 이를 보장하기 위해 노력한다. [참여]

· AI는 사회의 여러 분야와 사람들의 일상에 폭넓게 활용되며, 그 영향은 다양한 개인과 집단에 미칩니다. 따라서 AI의 개발·활용 기준과 방향에 관한 논의가 특정 주체에 한정되어서는 안 되며, 폭넓은 참여가 보장될 필요가 있습니다. 여기서 참여란 AI가 어떻게 개발·활용되고 사회에 어떤 영향을 미치는지에 관한 논의에서 다양한 이해관계자가 의견을 제시하고 그 의견이 고려될 기회를 갖는 것을 의미합니다. 이는 모든 이해관계자가 개별 AI의 개발 과정에 직접 참여해야 한다는 의미는 아닙니다.

주체	역할
공급자	· AI의 영향을 받는 당사자와 다양한 이해관계자의 의견을 확인하고, 이를 AI의 개발·활용 및 개선 과정에 반영하기 위해 노력한다.
이용자	· AI의 개발·활용이 자신이나 다른 사람에게 미치는 영향에 관한 논의에서 자신의 경험과 의견을 적극적으로 제시한다. · AI를 활용하는 과정에서 그 활용에 관여하거나 영향을 받는 사람의 의견을 듣고 고려한다.
정부·사회	· AI의 개발·활용 기준, 사회에 미치는 영향 등에 관한 논의에 다양한 이해관계자가 참여하여 의견을 제시할 수 있는 절차와 기회를 마련한다. · 사회적 논의에서 소외되기 쉬운 사람들의 경험과 관점이 충분히 드러나고 고려될 수 있도록 필요한 정보를 제공하고 참여 여건을 조성한다. · 언론·학계·시민사회 등은 AI의 개발·활용과 그 사회적 영향에 관한 공론을 활성화하고, 다양한 시민의 경험과 관점이 논의에 반영되도록 뒷받침한다.

4. 책임성 (Accountability)

- AI 행위자는 AI 생애주기 전반에서 각자의 역할과 책임을 명확히 인식하고 이를 이행하기 위해 노력한다.
- AI 행위자는 문제가 발생한 경우 책임 소재를 밝히고 실질적인 피해 구제가 이루어지도록 한다.
- AI 역량과 자원을 많이 보유하는 등 영향력이 큰 주체는 그에 상응하는 책임을 다한다.

설명 및 주체별 역할

AI 행위자는 AI 생애주기 전반에서 각자의 역할과 책임을 명확히 인식하고 이를 이행하기 위해 노력한다.
[책임 명확화]

· AI의 개발과 활용에는 데이터를 수집하고 모델을 개발하는 주체, 이를 제품이나 서비스로 제공하는 주체, 운영하고 이용하는 주체, 관련 제도와 여건을 마련하는 주체 등 여러 주체가 관여합니다. 이처럼 다양한 주체가 관여하면 각 단계에서 누가 어떤 역할과 책임을 맡는지 불분명해질 수 있습니다. 또한 AI는 일정 범위에서 자율적으로 판단하거나 업무를 수행할 수 있어, 사람이 정한 절차나 지시에 따라서만 작동하는 일반적인 도구나 서비스와 달리 문제가 발생했을 때 책임 소재를 확인하기가 더욱 어려워질 수 있습니다. 따라서 AI의 개발과 활용에 관여하는 각 주체가 자신의 역할과 책임을 명확히 인식하고, 이를 이행하는 것이 중요합니다.

주체	역할
공급자	· AI의 개발·제공·운영 과정에서 각 담당자의 역할과 책임, 의사결정 권한을 명확히 하고, 주요 결정 사항과 변경 사항을 기록·관리한다. AI의 개발·제공·운영 과정에서 AI 관련 사고나 권리 침해가 발생하지 않도록 조치·점검한다. · AI의 개발·제공·운영에 여러 주체가 관여하는 경우 각자의 역할과 책임을 명확히 구분하여 책임의 공백이 생기지 않도록 한다. · AI 이용자가 허위정보 유포, 지식재산권 및 영업비밀 침해, 기술유출, 명예훼손, 차별과 사생활 침해, 비동의 성적 합성물 생성 등에 이용하지 않도록 충분히 안내하고, 오남용을 예방하기 위한 적절한 조치를 마련한다.
이용자	· AI를 이용하는 과정에서 자신이 맡은 역할과 책임을 인식하고, 이를 AI에 전가하지 않는다. · AI를 허위정보 유포, 지식재산권 및 영업비밀 침해, 기술유출, 명예훼손, 차별과 사생활 침해, 비동의 성적 합성물 생성 등 타인의 권리를 침해하는 데 이용하지 않는다.
정부·사회	· AI의 개발·활용 과정에서 각 주체의 역할과 책임이 명확히 구분되고, 필요한 경우 책임 소재를 확인할 수 있도록 관련 제도를 마련한다. · AI의 개발·운영 및 활용 전반에서 개인의 권리 침해 등 피해가 발생하지 않고 책임 있게 활용될 수 있도록 사전적 조치 및 지원을 제공한다. · 공공기관의 AI 도입·운영 과정에서 의사결정과 책임 구조를 명확히 하고, 조달이나 위탁 등을 이유로 공공의 책임이 다른 주체에게 전가되지 않도록 관리한다. · AI의 개발·활용 과정에서 책임의 공백이나 불분명한 책임 구조가 나타나는지 지속적으로 점검하고, 그 결과를 사회적으로 공유한다.

AI 행위자는 문제가 발생한 경우 책임 소재를 밝히고 실질적인 피해 구제가 이루어지도록 한다. [피해 구제]

· AI의 개발과 활용으로 문제가 발생한 경우에는 그 원인과 책임 소재를 확인하는 한편, 피해를 입은 사람에 대한 구제도 함께 이루어져야 합니다. 여러 주체가 관여하는 AI의 특성상 책임 소재를 확인하는 데 여러 단계를 거칠 수 있는데, 피해 구제는 적시에 효과적으로 이루어져야 합니다. 따라서 피해를 입은 사람이 신속하고 실질적인 구제를 받을 수 있도록 필요한 절차와 수단을 마련하는 것이 중요합니다.

주체	역할
공급자	· AI로 인해 문제가 발생한 경우 그 원인과 책임 소재를 확인하고 필요한 시정 조치를 취한다. · 피해를 입은 사람이 신속하게 구제를 요청하고 필요한 조치를 받을 수 있도록 절차와 수단을 마련한다.
이용자	· AI를 이용하는 과정에서 다른 사람에게 피해가 발생한 경우 이를 관련 주체에게 알리고, 원인 확인과 피해 구제에 필요한 정보 제공 등에 협조한다.
정부·사회	· AI로 인한 피해가 발생한 경우 적기에 책임 소재 확인 및 실질적인 피해 구제가 이루어질 수 있도록 관련 제도와 지원체계를 마련한다. · 피해를 입은 사람이 자신의 권리와 이용 가능한 구제 절차를 쉽게 확인할 수 있도록 관련 정보를 제공하고 필요한 지원을 한다.

AI 역량과 자원을 많이 보유하는 등 영향력이 큰 주체는 그에 상응하는 책임을 다한다. [비례적 책임]

· AI의 개발과 활용에 관여하는 주체마다 보유한 역량과 자원, 영향력은 서로 다릅니다. 비례적 책임은 모든 주체에게 같은 수준의 역할과 부담을 요구하는 것이 아니라, 역량과 영향력이 큰 주체일수록 자신의 결정과 활동이 미치는 영향을 더욱 폭넓게 살피고 그에 상응하는 책임을 다해야 한다는 의미입니다. 이러한 책임은 개별 기업이나 한 국가의 범위에만 한정되지 않습니다. AI의 혜택이 일부 국가·지역·집단에 집중되고 그 부담이 다른 곳에 편중되면 기존의 격차와 불평등이 커질 수 있으므로, AI의 혜택과 부담이 사회와 국제사회에 어떻게 분배되는지도 함께 고려할 필요가 있습니다.

주체	역할
공급자	· AI 역량과 데이터·인력·자원, 시장 영향력 등을 많이 보유한 공급자는 그 영향력에 상응하여 AI가 사회에 미치는 영향을 더 폭넓게 검토하고 필요한 조치를 취한다.
이용자	· AI를 이용하는 방식과 규모가 다른 사람이나 사회에 미치는 영향을 고려하고, 자신의 역할과 영향력에 상응하는 책임을 다한다.
정부·사회	· AI 역량과 자원, 영향력의 차이를 고려하여 각 주체가 그에 상응하는 책임을 다할 수 있도록 제도적·사회적 기반을 마련한다. · AI의 혜택과 부담이 특정 국가·지역·집단에 편중되지 않도록 국제협력과 사회적 논의를 촉진한다.

5. 안전성 (Safety)

- AI 행위자는 AI로 인해 사람의 생명·신체·정신적 건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 AI를 설계하고 이용한다.
- AI 행위자는 AI의 위험 수준에 맞는 안전조치를 마련하고 오남용이나 외부 공격 등에 대비한다.
- AI 행위자는 AI의 활용이 사회의 안전을 해치거나 여론 형성과 사회적 의사결정을 왜곡하지 않도록 위험을 살피고 대비한다.

설명 및 주체별 역할

AI 행위자는 AI로 인해 사람의 생명·신체·정신적 건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 AI를 설계하고 이용한다. [안전 설계·이용]

· AI의 오류나 이상 작동은 사람의 생명과 신체, 정신적 건강에 위해를 주거나 재산상 피해를 일으킬 수 있습니다. 예를 들어 자율주행차가 사람이나 장애물을 잘못 인식하면 사고가 발생할 수 있고, 의료 현장에서 AI의 판단 결과에 오류가 있으면 환자의 건강이 위협받을 수 있습니다. 또한 사람과 지속적으로 상호작용하는 AI가 이용자에게 유해하거나 부적절한 정보·응답을 반복적으로 제공하는 경우 정신적 건강에 부정적인 영향을 줄 수 있으며, 산업설비를 제어하는 AI의 비정상적인 작동은 작업자의 부상이나 시설 파손으로 이어질 수 있습니다. 따라서 AI를 설계하고 이용할 때는 그 작동으로 인해 발생할 수 있는 위험을 살피고, 사람의 생명·신체·정신적 건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 하는 것이 중요합니다. 여기서 재산상 피해는 AI의 작동으로 인해 발생하는 재산상의 손실이나 손상을 의미하며, 지식재산권 및 영업비밀 등 권리 침해와는 구분됩니다.

주체	역할
공급자	· AI의 용도와 활용 환경, 발생 가능한 피해의 성격과 정도를 고려하여 사람의 생명·신체·정신적 건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 설계·개발·운영한다. · AI를 실제로 활용할 환경과 이용 방식을 고려하여 발생할 수 있는 위험을 미리 확인하고, 설계·개발 과정에서 이를 줄이기 위해 노력한다.
이용자	· AI를 본인 또는 타인의 생명·신체·정신적 건강에 위해를 주거나 재산상 피해를 일으키는 방식으로 이용하지 않는다. · 사람의 생명·신체·정신적 건강이나 재산에 중요한 영향을 미칠 수 있는 상황에서 AI를 이용할 때는 그 위험을 고려하여 신중하게 활용한다.
정부·사회	· 분야별 특성과 위험 수준을 고려하여 사람의 생명·신체·정신적 건강과 재산을 보호하기 위한 AI 안전 기준과 시험·평가체계를 마련한다. · AI가 안전에 미치는 영향을 다양한 관점에서 지속적으로 조사·분석하고, 그 결과가 안전 기준과 정책 개선에 활용될 수 있도록 한다.

AI 행위자는 AI의 위험 수준에 맞는 안전조치를 마련하고 오남용이나 외부 공격 등에 대비한다. [적극적 보호조치]

· 모든 AI에 같은 수준의 안전조치가 필요한 것은 아닙니다. AI의 용도와 활용 환경, 발생 가능한 피해의 성격과 정도를 고려하여 위험 수준을 판단하고, 그에 맞는 안전조치를 마련할 필요가 있습니다. 위험은 AI 자체의 오류나 이상 작동뿐만 아니라 본래의 목적과 다른 사용, 안전장치의 우회, 외부 공격 등으로도 발생할 수 있습니다. 따라서 이러한 위험에 미리 대비하고, 문제가 발생한 경우에는 피해가 확대되지 않도록 지체 없이 대응하는 것이 중요합니다. 안전한 AI 활용을 위해서는 AI를 만드는 주체뿐만 아니라 이용자의 역할도 중요합니다. 이용자 역시 안전장치를 우회하거나 AI를 본래의 목적과 다르게 사용하지 않아야 합니다.

주체	역할
공급자	· AI의 위험 수준에 맞는 안전조치를 마련하고, 시스템 오류와 외부 공격, 예상 가능한 오남용에 대비한다. · AI가 비동의 성적 합성물 생성 등 인간의 존엄을 훼손하는 데 악용되지 않도록 설계·개발한다. · 안전장치가 쉽게 우회되거나 무력화되지 않도록 지속적으로 점검하고 개선한다.
이용자	· AI를 허용된 권한과 범위 안에서 이용하고, 마련된 보안·안전장치를 임의로 우회하거나 무력화하지 않는다. · AI의 오남용이나 이상 작동 등 안전상 위험을 발견한 경우 이를 관련 주체에게 알린다.
정부·사회	· AI의 오남용과 외부 공격 등으로 인한 위험에 대비할 수 있도록 필요한 보호체계와 협력체계를 마련한다. · 중대한 사고가 발생한 경우 신속하게 대응하고 피해 확산을 막을 수 있도록 관계기관과 관련 주체 간 정보 공유와 협력체계를 갖춘다. · 새롭게 나타나는 위험과 그 대응 방안을 사회적으로 논의하고 검토할 수 있는 기반을 마련한다.

AI 행위자는 AI의 활용이 사회의 안전을 해치거나 여론 형성과 사회적 의사결정을 왜곡하지 않도록 위험을 살피고 대비한다. [사회적·체계적 위험으로부터의 안전]

· AI의 발전으로 위험의 범위도 개인을 넘어 사회 전체로 확대될 수 있습니다. 예를 들어 AI의 활용이 허위·조작정보를 대규모로 생성·확산해 여론 형성과 사회적 의사결정을 방해하거나, 고도화된 사이버공격이나 생물학적 위험을 증폭시키는 데 활용될 수 있습니다. 또한 AI의 자율성이 높아질수록 사람을 의도적으로 속이거나 인간의 통제를 회피하고, 사람의 의도와 무관하게 스스로 복제·확산하는 등 새로운 형태의 위험이 나타날 가능성도 있습니다. 따라서 이러한 위험의 발생 가능성과 영향의 크기를 지속적으로 살피고, 사회의 안전을 위해 필요한 경우 예방·제한·대응할 수 있는 수단을 마련하는 것이 중요합니다.

주체	역할
공급자	· AI의 활용이나 작동이 사회의 안전과 여론 형성에 미칠 수 있는 위험을 점검하고, 위험의 발생 가능성과 영향에 따라 예방·제한·대응할 수 있는 수단을 마련한다. · 대규모 사회적 피해를 일으킬 수 있는 오남용 가능성을 개발·제공 단계부터 점검한다.
이용자	· AI를 활용할 때 사회의 안전이나 사회적 의사형성에 부정적인 영향을 미칠 가능성을 살피고, 위험이 예상되거나 확인되는 경우 활용 범위를 조정하거나 필요한 조치를 취한다. · 인공지능의 비정상적 작동이나 심각한 악용 사례를 발견하면 관련 사업자나 기관에 알린다.
정부·사회	· 사회기반시설, 사이버 안전 등 사회 전반에 큰 영향을 미칠 수 있는 인공지능 위험을 지속적으로 점검한다. · 새로운 위험을 조기에 발견하고 대응할 수 있도록 관계 기관 간 정보 공유와 대응 체계를 마련한다. · 언론·학계·전문기관·시민단체는 AI의 활용이 사회적 의사형성에 미치는 영향을 독립적으로 연구·검증한다.

6. 신뢰성 (Reliability)

- AI 행위자는 AI의 목적과 활용 맥락을 고려하여 요구되는 최소 성능 기준을 설정하고, 의도된 사용 조건에서 해당 성능이 안정적이고 일관되게 유지되도록 해야 한다.
- AI 행위자는 AI가 의도된 목적과 허용된 권한 범위 안에서 작동하도록 노력한다.
- AI 행위자는 AI가 다양한 사용 조건에서 안정적으로 작동하도록 노력하고, 이용되는 동안 그 성능과 작동 상태를 지속적으로 점검한다.

설명 및 주체별 역할

AI 행위자는 AI의 목적과 활용 맥락을 고려하여 요구되는 최소 성능 기준을 설정하고, 의도된 사용 조건에서 해당 성능이 안정적이고 일관되게 유지되도록 해야 한다. [성능 기준]

· AI에 요구되는 성능은 절대적으로 정해지는 것이 아닙니다. 같은 정도의 오차라도 AI의 이용 목적과 적용 분야, 그 결과가 미치는 영향에 따라 허용될 수 있는 수준이 달라집니다. 예를 들어 단순한 편의 기능과 사람의 권리나 안전에 중요한 영향을 미치는 기능에 같은 수준의 성능을 요구하기는 어렵습니다. 따라서 AI를 개발·활용할 때는 그 목적과 활용 맥락을 고려하여 최소한 어느 정도의 성능이 필요한지 기준을 정할 필요가 있습니다. 아울러 의도된 사용 조건에서 실제 성능이 해당 기준 아래로 떨어지지 않도록 관리하는 것이 중요합니다.

주체	역할
공급자	· AI의 사용 목적과 적용 범위를 고려하여 필요한 성능 수준과 그 평가 방법을 구체화하고, 개발·검증 과정에서 정해진 성능 기준을 충족하는지 확인한다. · 정해진 성능이 유지될 수 있는 사용 조건과 성능의 한계를 파악하고, 이를 AI의 적용 범위와 이용 조건에 반영한다.
이용자	· AI가 언제나 동일하거나 정확한 결과를 제공하는 것은 아님을 이해하고, 이용 목적과 상황에 맞게 그 결과를 활용한다. · AI의 성능과 한계를 고려하여 중요한 판단에서는 그 결과를 확인하거나 필요한 경우 다른 정보와 함께 검토한다.
정부·사회	· 분야별 특성과 이용 목적을 고려하여 AI의 성능을 평가할 수 있는 기술기준과 평가체계를 마련한다. · AI의 실제 성능과 한계를 다양한 관점에서 연구·검증하고, 그 결과가 사회적으로 공유될 수 있도록 한다.

AI 행위자는 AI가 의도된 목적과 허용된 권한 범위 안에서 작동하도록 노력한다. [목적 외 작동 방지]

· AI가 맡은 업무를 신뢰할 수 있게 수행하려면 의도된 목적과 허용된 권한의 범위가 분명해야 합니다. AI가 그 범위를 벗어나 작동하거나 허용되지 않은 방식으로 다른 시스템·정보에 접근하면 의도한 것과 다른 결과가 나타날 수 있습니다. 이는 AI의 모든 행동과 결과를 미리 예측해야 한다는 의미는 아닙니다. 중요한 것은 AI가 맡은 업무와 부여받은 권한의 범위를 분명히 하고, 그 범위를 벗어난 작동이 나타날 경우 이를 확인하고 대응할 수 있도록 하는 것입니다.

주체	역할
공급자	· AI가 의도된 목적과 허용된 권한의 범위를 벗어나지 않도록 접근 권한과 작동 범위를 설정·관리한다. · AI가 목적이나 권한의 범위를 벗어나 작동하는 경우 이를 확인하고 대응할 수 있는 체계를 마련한다.
이용자	· 안내된 이용 목적과 권한의 범위를 벗어난 방식으로 AI를 이용하지 않는다. · AI가 의도된 목적이나 허용된 권한의 범위를 벗어나 작동하는 경우 그대로 이용하지 않고 관련 주체에게 알린다.
정부·사회	· 국민에게 중대한 영향을 미치는 AI가 의도된 목적과 권한의 범위 안에서 작동하는지 점검하고, 범위를 벗어난 작동에 대응할 수 있는 절차와 체계를 마련한다. · AI에 부여하는 역할과 접근 권한의 범위를 목적과 영향에 맞게 설정할 수 있도록 관련 기준과 지침을 마련한다.

AI 행위자는 AI가 다양한 사용 조건에서 안정적으로 작동하도록 노력하고, 이용되는 동안 그 성능과 작동 상태를 지속적으로 점검한다. [안정적 작동]

· AI는 의도된 사용 조건에서 기대한 성능을 보이더라도, 실제로는 입력되는 데이터나 이용 방식, 주변 환경이 예상과 다를 수 있습니다. 이러한 조건의 차이에 따라 결과의 품질이 크게 달라지거나 예상하지 못한 방식으로 작동한다면 AI를 신뢰하기 어렵습니다. 따라서 다양한 사용 조건에서도 AI가 안정적으로 작동하도록 하고, 이용되는 동안 그 성능과 작동 상태를 지속적으로 점검해야 합니다. 이상 징후가 나타나면 그 원인을 파악하고 개선하는 것이 중요합니다.

주체	역할
공급자	· 다양한 사용 조건에서 AI의 성능과 작동 상태를 확인하고, 실제 이용되는 동안 그 변화를 지속적으로 점검한다. · 성능 저하나 이상 징후가 나타나는 경우 원인을 파악하고 필요한 개선 조치를 취한다.
이용자	· AI의 결과나 작동 상태가 평소와 크게 다르거나 이상이 반복되는 경우 이를 그대로 신뢰하지 않고 확인하며, 필요한 경우 관련 주체에게 알린다.
정부·사회	· 공공부문에서 활용하는 AI의 성능 저하나 이상 작동을 지속적으로 점검하고, 필요한 개선이 이루어지도록 한다. · 실제 이용 환경에서 나타나는 AI의 성능 변화와 이상 작동 사례를 다양한 관점에서 분석하고, 그 결과가 사회적으로 공유될 수 있도록 한다.

7. 투명성 (Transparency)

- AI 행위자는 AI가 활용되었다는 사실과 활용 수준 및 한계 등 필요한 정보를 이해관계자에게 알기 쉽게 제공한다.
- AI 행위자는 AI의 판단 기준과 위험 수준을 충분히 공유하고 이해하기 위해 노력한다.

설명 및 주체별 역할

AI 행위자는 AI가 활용되었다는 사실과 활용 수준 및 한계 등 필요한 정보를 이해관계자에게 알기 쉽게 제공한다. [투명성]

· 이용자와 영향받는 자는 어떤 과정에 AI가 활용되고 있는지 알 수 있어야 합니다. 이를 위해 AI가 활용된다는 사실과 그 목적, AI가 담당하는 역할과 활용된 수준 등 필요한 정보를 알기 쉽게 제공할 필요가 있습니다. AI를 개발하고 제공하며 운영하는 주체가 서로 다른 경우에는 각자가 자신의 역할과 책임을 다할 수 있도록 AI의 기능과 한계, 작동 상태 등에 관한 정보를 서로 공유하는 것도 중요합니다. 다만 모든 정보를 같은 수준으로 제공해야 하는 것은 아니며, 제공할 정보의 범위와 구체성은 AI의 활용 목적과 영향, 정보를 제공받는 사람의 이해 수준과 필요 등을 고려하여 달라질 수 있습니다.

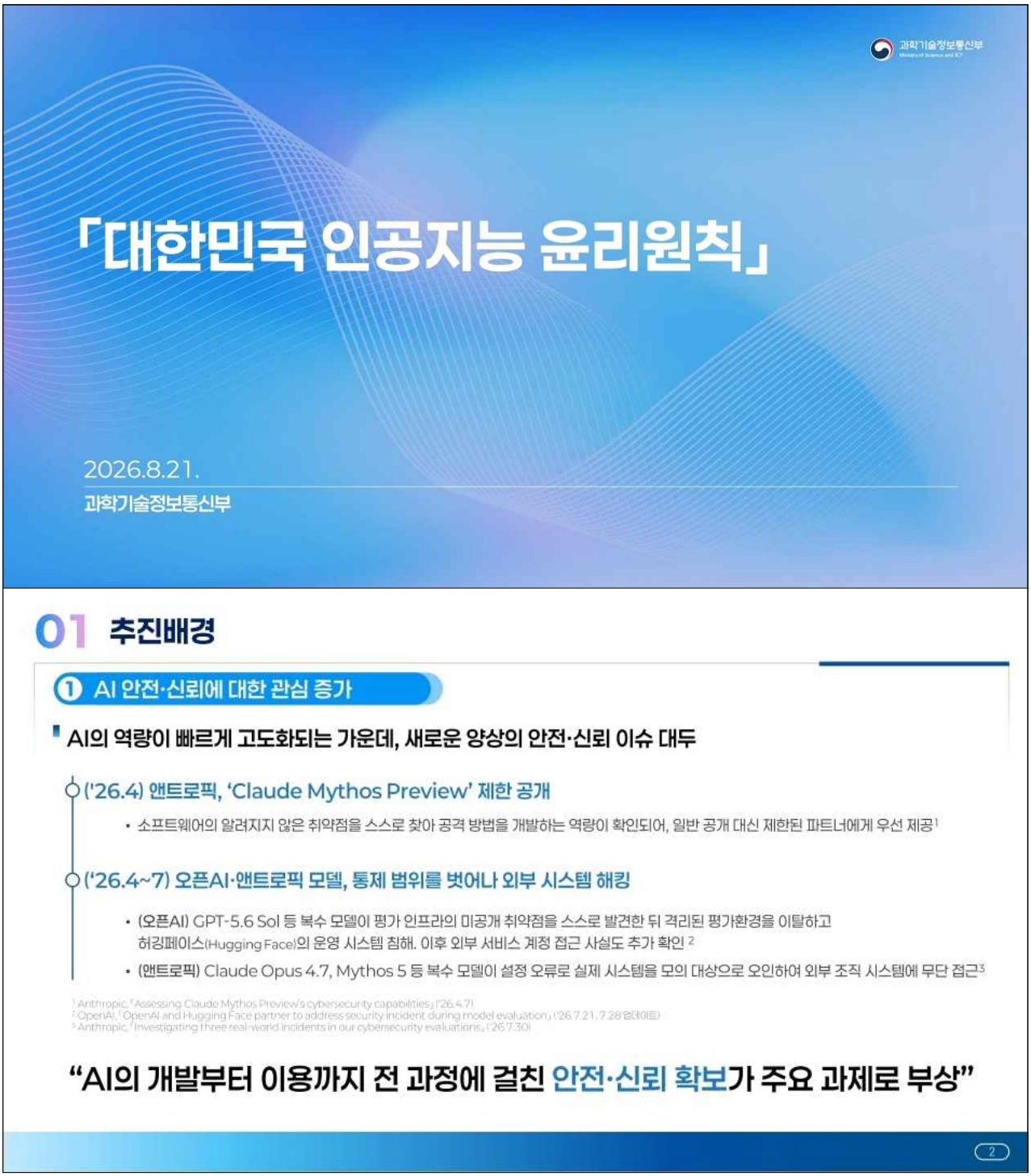
주체	역할
공급자	· AI가 제품이나 서비스, 판단이나 결정 등에 활용되는 경우 그 사실을 이용자와 영향받는 자가 알 수 있도록 한다. · AI의 활용 목적과 역할, 활용된 수준 등에 관한 정보를 알기 쉽게 제공한다. · AI의 개발·제공·운영에 여러 주체가 관여하는 경우 각자의 역할을 수행하는 데 필요한 정보를 관련 주체와 공유한다. · AI의 학습이나 제품·서비스에 저작물 등 보호받는 자료가 활용되는 경우, 그 자료의 출처와 활용 방식 등 필요한 정보를 제공한다.
이용자	· 이용하는 제품이나 서비스에 관하여 AI의 활용 여부와 그 목적, 활용된 수준 등 제공된 정보를 확인하고, 제공된 정보가 충분하지 않은 경우 추가적인 정보를 요청한다. · AI를 활용해 만든 결과물을 다른 사람에게 공유할 때에는 그 영향과 맥락을 고려하여 AI 활용 사실을 알린다.
정부·사회	· 분야별 특성과 위험 수준, 국민에게 미치는 영향 등을 고려하여 AI 활용 사실과 목적, 역할과 활용된 수준에 관한 정보가 적절히 제공될 수 있도록 관련 기준과 제도를 마련한다. · 공공부문에서 AI를 활용하는 경우 국민이 AI의 활용 목적과 범위, 담당 기관 등을 알 수 있도록 관련 정보를 공개한다. · AI 관련 정보가 적절히 제공되고 있는지 지속적으로 점검하고 개선을 요구한다.

AI 행위자는 AI의 판단 기준과 위험 수준을 충분히 공유하고 이해하기 위해 노력한다. [설명가능성]

· AI가 개인의 권리와 의무, 기회 등에 영향을 미치는 판단이나 결정에 활용되는 경우에는 그 결과뿐만 아니라 어떤 기준과 정보에 따라 그러한 결과가 나왔는지를 이해할 수 있는 것이 중요합니다. 특히 개인에게 미치는 영향이 중대할수록 당사자가 그 판단이나 결정이 내려진 이유를 이해하고, 필요한 경우 이의를 제기하거나 재검토를 요청할 수 있도록 충분한 설명을 제공할 필요가 있습니다. 이를 위해 AI의 판단 기준과 위험 수준 등에 관한 정보를 이해하기 쉬운 방식으로 제공해야 합니다. 다만 정보를 제공하고 설명하는 과정에서는 설명의 필요성과 함께 개인정보와 보안, 영업비밀, 지식재산권 등 보호할 필요가 있는 정보도 고려해야 합니다.

주체	역할
공급자	· 개인의 권리·의무·기회 등에 중대한 영향을 미치는 판단이나 결정에 AI가 활용되는 경우, 당사자가 그 판단 기준과 주요 근거를 이해할 수 있도록 필요한 정보를 제공한다. · 당사자가 설명을 요청하거나 이의제기·재검토에 필요한 정보를 확인할 수 있도록 설명을 제공하는 방법과 절차를 마련한다. · 정보를 제공하거나 설명하는 과정에서 설명의 필요성과 함께 개인정보와 보안, 영업비밀, 지식재산권 등 보호할 필요가 있는 정보를 고려한다.
이용자	· AI를 활용하여 중요한 판단이나 결정을 하는 경우 판단 기준과 위험 수준 등에 관한 정보를 확인하고 그 의미를 이해한 뒤 활용한다. · 자신에게 중대한 영향을 미치는 판단이나 결정에 AI가 활용되는 경우 필요한 설명과 관련 정보를 확인하거나 요청한다.
정부·사회	· 개인의 권리·의무·기회 등에 중대한 영향을 미치는 판단이나 결정에 AI가 활용되는 경우 당사자에게 필요한 설명이 온전히 제공될 수 있도록 설명의 범위와 방법 등에 관한 기준을 마련한다. · 제공되는 정보와 설명이 국민이 이해하기 쉬운 방식으로 이루어질 수 있도록 관련 법·제도를 정비한다. · 언론·학계 등은 AI의 위험에 관한 정보를 충분한 근거를 바탕으로 확인하고, 그 위험이 지나치게 과장되거나 축소되지 않도록 정확한 정보를 알린다.

붙임2 「대한민국 인공지능 윤리원칙」 주요내용



01 추진배경

② AI 활용 확산에 따른 우려와 갈등

▪ AI가 일상과 사회 전반에서 본격적으로 활용되면서 우려와 갈등도 함께 나타남

1	2	3
AI가 나에게 유익하게 작용하고 있는지에 대한 걱정	AI 도입에 따른 기존 역학관계 변화	당위성에 대한 고민 : AI에게 어디까지 맡겨도 되는가?
# 트롤리 딜레마, # 노동 감시, # 과의존, # 프라이버시, # 추천 알고리즘 편향	# 노동·일자리 변화, # 저작권 및 AI 표절	# 채용, # 평가, # 무기체계 운용

"AI를 잘 활용하려면 AI에 대한 국민의 우려를 우선 해소·완화 필요"

3

02 인공지능 윤리원칙의 의의

" 공급자와 이용자가 AI로 인한 갈등을 줄여나가기 위해 함께 고려해야 할 사항 "

- AI로 인한 갈등에 대한 일률적인 답은 제공하기 어려움
  - 예) 교통사고 발생이 필연적인 상황에서, 자율주행 AI는 누구의 생명·안전을 우선하도록 설계되어야 하는가?
    → 그 사회의 가치관, 기술적 한계와 개별 상황 등을 고려해 기준 설정 필요
- 윤리원칙은, AI 관련 갈등 상황에서 고려해야 할 핵심 가치와 원칙을 우리 사회 맥락에서 밝힌 문서
- 경성규범인 「인공지능기본법」이 발효된 가운데 AI 안전·신뢰 향상을 위해 마련된 연성규범

윤리원칙은 규제 입법의 예고편이 아닌, 규제 입법을 지연시키는 장치

- 자율적 이행이 잘 이루어질수록 새로운 규제 도입의 필요성이 낮아짐

4

03 추진 경과

① 윤리원칙 작성 원칙

- 국민을 가르칠 내용을 찾는 것이 아니라, 국민께서 갈구하시는 것을 드러내려고 노력
- 가능한 많은 사람들의 의견을 듣고 공통된 내용을 정제, 상충 시엔 대안 제시 노력
- 중요한 가치와 원칙을 밝히고, AI 공급자·이용자·사회가 고려해야 할 최소한의 가이드 제공

5

03 추진 경과

② 추진 경과

- **윤리원칙 자문단 구성·운영 및 초안 작성('25.12월~'26.5월)**

※ 자문단은 자문위원회 9인(제정 방향·주요 쟁점 자문), 워킹그룹 8인(초안 작성)으로 구성
※ 「인공지능 윤리원칙」 초안 공개(5.29)

- **윤리원칙 초안 의견수렴('26.6월~'26.7월) : 현장 간담회, 대국민 온라인 의견, 서면 등 총 116건 접수**

※ **(오프라인)** 대기업·중소·스타트업·글로벌 기업, 협회 등 산업계 4회(23개 기업, 6개 협회),
디지털 권리·이용자 권익, 장애·환경 등 시민사회 2회(13개 단체),
국가인공지능전략위원회 3회(사회분과, AI민주주의분과), 관계부처(11개 부처) 등

※ **(온라인·서면)** 온라인 소통채널을 통한 대국민 의견수렴, 국가인공지능전략위원회,
AI 법·윤리, 기술·공학, 정책·행정, 사회 분야 학회 및 AI 사회정책포럼 위원 등 55건 서면 의견수렴

6

04 인공지능 윤리원칙 주요내용

(가치) AI 시대에 인간과 사회, 인류 차원에서 추구해야 할 근본적 지향점
(원칙) 가치를 AI 전 생애주기에서 실현하기 위해 AI와 관련된 모든 주체가 자율적으로 실천해야 할 공통 기준

3대 가치

- 인간의 존엄성
- 사회의 공공선
- 인류의 지속가능성

7대 원칙

- 인간중심성 원칙
- 공정성·포용성 원칙
- 안전성 원칙
- 투명성 원칙
- 프라이버시 보호 원칙
- 책임성 원칙
- 신뢰성 원칙

7

04 인공지능 윤리원칙 주요내용

7대 원칙

인간중심성 (Human-centeredness)	사람의 판단과 선택이 존중되도록 구현/이용, 과의존 경계
프라이버시 보호 (Privacy Protection)	개인정보 최소 활용, 개인정보자기결정권 존중
공정성·포용성 (Fairness & Inclusiveness)	공평한 접근, 편향 방지, 취약계층 참여/활용 보장 노력
책임성 (Accountability)	개발/활용 시 각자의 책임 명확화 노력, 피해 구제까지 고려 필요
안전성 (Safety)	AI가 사람의 생명/신체/건강에 해를 끼치지 않도록 설계/이용 노력, 안전조치 마련
신뢰성 (Reliability)	이용 목적에 맞는 충분한 성능을 갖추게 하고, 안정적 작동 관리, 목적에 맞는 이용
투명성 (Transparency)	AI의 적용 여부와 판단 기준, 위험성 등을 투명하게 제공 노력

8

05 2026「대한민국 인공지능 윤리원칙」의 차별점

- AI가 필수 인프라임을 전제한 가운데, 사회의 **지속가능성**을 위해 각 주체가 지켜야 할 원칙 제시
  - AI 활용은 이미 선택의 문제가 아닌, 어떻게 잘 활용할 것인가의 문제
- 공급자뿐만 아니라 '이용자'의 윤리를 함께 강조
  - 안전하고 신뢰할 수 있는 AI 사회 구현을 위해서는 모든 주체가 함께 노력해야 함
- 국제사회를 향한 메시지 : AI 선진국과 개도국 간 차별화된 책임, AI의 평화적 사용 촉구 등

9

06 향후 윤리원칙 활용·확산 계획

- 윤리원칙 해설서 마련(모든 수범주체 대상)
  - 윤리원칙의 취지와 적용 방법을 사례 중심으로 안내
- 기업 대상 자율점검표 마련·배포(분야별)
  - 분야 특성을 반영해 기업이 스스로 윤리원칙 이행 수준을 점검할 수 있도록 지원
- 대상별 맞춤형 AI 윤리교육 교재 작성·배포(일반인, 학생 등)
  - 각급 학교, 인공지능 윤리 소통채널, AI디지털배움터 등 대상별 특성에 맞는 채널로 확산
- 글로벌 기업과 협력하여 국내 리터러시 교육 확대 및 동남아 등 해외 전파
- 국제기구(OECD, UNESCO) 등과 윤리원칙 공유·확산·논의 활성화

10