

인공지능 윤리원칙 2차안

I. 서문

1. AI 시대의 기회와 과제

인공지능(AI)은 생산성을 높이고, 새로운 산업과 서비스를 만들며, 국민의 삶을 더 편리하고 풍요롭게 하는 기술이다. 또한 기후 변화와 재난 대응, 고령화, 지역 격차 등 우리 사회가 함께 풀어야 할 문제를 해결하는 데에도 기여할 수 있다. 최근 AI는 생성형 AI를 넘어 스스로 과업을 계획하고 수행하는 에이전틱 AI와 물리적 환경에서 작동하는 피지컬 AI로 발전하고 있다. 이러한 발전은 인간의 창의성과 문제해결 역량을 넓히는 한편 AI의 활용 범위를 빠르게 확장하고 있다.

AI가 사회 전반에 폭넓게 활용되면서 새로운 과제도 나타나고 있다. 군사·안보 목적의 활용과 이에 대한 인간의 통제 범위, 창작과 표절의 경계, 학습에 활용된 데이터와 전문 지식에 대한 정당한 보상, 일터에서의 감시와 일자리 구조 변화, 허위·조작 정보의 확산과 정보 편중에 따른 여론 다양성 저하 등이 그러한 예이다.

AI는 정보와 콘텐츠가 추천·생성·확산되는 방식을 바꾸어 여론 형성과 시민의 판단, 민주적 의사 결정에 영향을 미칠 수 있다. 따라서 AI의 개발과 활용은 이러한 영향을 고려하여 의견의 다양성과 자유로운 소통을 존중하고, 시민의 참여를 넓히는 방향으로 이루어져야 한다. 또한 성별, 장애, 연령, 지역, 인종, 사회적 신분 등에 관계없이 모든 시민을 포함한 다양한 이해관계자가 AI의 개발과 활용을 함께 논의할 수 있는 기회가 마련되어야 한다. 이러한 숙의의 과정은 서로 다른 가치와 관점을 조정하고, 사회적 신뢰와 민주주의의 발전을 뒷받침한다.

AI가 가져올 기회를 살리고 새로운 과제에 대응하기 위해서는, AI의 개발과 활용이 인간의 존엄성을 바탕으로 사회의 공공선을 추구하고 인류의 지속가능성에 기여하는 방향으로 이루어져야 한다. 그 과정에서 헌법상 기본권과 국제인권규범에 따른 인권이 보장되어야 한다.

2. 본 원칙의 위상과 성격

이러한 지향을 실천하기 위해 본 윤리원칙은 AI와 관계되는 모든 주체가 함께 참고할 수 있는 공통의 가치와 원칙을 제시한다. 윤리원칙은 특정 기술이나 분야에 한정된 세부 기준이 아니라 범정부적·범사회적으로 적용되는 상위 원칙이다.

본 윤리원칙은 법령을 대체하지 않으며 분야별 윤리기준과 가이드라인의 바탕이 된다. 법령이 구체적인 의무와 기준을 정한다면, 윤리원칙은 빠르게 변화하는 환경에서 각 주체가 바람직한 방향을 스스로 판단하고 실천하도록 돕는다.

본 윤리원칙은 법적 구속력이 없는 자율적 원칙이다. 각 주체는 AI의 개발·활용 목적과 영향, 자신의 역할과 역량을 고려하여 윤리원칙을 실천한다.

3. 본 원칙의 적용과 공동의 실천

윤리원칙의 실천에는 여러 주체의 참여와 협력이 필요하다. 본 윤리원칙은 AI를 개발하고 제공하는 주체의 역할과 함께 이를 이용하는 주체의 역할도 다룬다. AI가 사회에 미치는 영향은 개발과 제공 단계에서만 정해지지 않으며, 실제로 어떤 목적과 방식으로 이용되는지에 따라 달라지기 때문이다.

개발자와 사업자는 신뢰할 수 있는 AI를 만들고 제공한다. 시민은 AI를 책임 있게 이용하고, AI의 개발과 활용에 관한 사회적 논의에 주체적으로 참여한다. 정부와 공공기관은 AI를 활용할 때 스스로 윤리원칙을 앞서 실천하며, 각 주체가 이를 실천할 수 있는 여건을 조성한다. 언론·학계·시민사회 등은 AI가 사회에 미치는 영향을 독립적으로 살피고 필요한 문제를 제기한다.

특히 시민은 데이터와 콘텐츠의 생성·제공과 공론장 참여를 통해 AI의 미래를 함께 만들어 가는 주체다. 각 주체는 시민의 이러한 역할을 존중하고 뒷받침한다.

윤리원칙을 실제로 적용하는 과정에서 가치와 원칙이 서로 긴장하거나 충돌할 수 있다. 본 원칙은 고정된 우선순위를 미리 제시하지 않는다. 대신 각 주체가 사안의 맥락과 영향을 살피고, 다양한 이해관계자의 논의를 거쳐 적절한 균형점과 실천 방안을 찾기를 권장한다. 다만 생명·안전이나 기본권에 중대한 영향을 미치거나 회복하기 어려운 피해가 예상되는 경우에는 피해를 예방하고 줄이는 것을 우선한다.

이러한 실천은 국내에만 머물러서는 안 된다. AI의 개발과 활용은 국제 평화와 인류 공동의 이익에 기여해야 한다. 특히 AI 관련 역량과 자원을 많이 보유한 국가와 기업은 그 영향력에 상응하는 책임을 다하고, AI의 혜택이 특정 국가·지역·집단에 집중되지 않도록 노력해야 한다. 이를 위해 우리나라는 안전하고 신뢰할 수 있는 AI의 발전과 책임 있는 활용을 촉진하고 국제사회와 협력한다.

4. 이행 지원과 지속적 점검

정부는 본 윤리원칙이 선언에 그치지 않고 우리 사회에서 실질적으로 이행되도록 책임 있게 뒷받침한다. 교육을 통해 사회 구성원이 AI 윤리의 가치와 원칙을 이해하고 실천하도록 지원하며, 각 주체의 자율적 이행에 필요한 정책과 제도를 마련한다.

본 윤리원칙은 한 번의 제정으로 완결되지 않는다. 정부는 기술과 사회의 변화에 따라 새롭게 제기되는 쟁점과 각 주체의 이행 경험을 바탕으로 윤리원칙을 정기적으로 점검하고 보완한다. 또한 다양한 주체의 의견을 주기적으로 수렴하고 이행 과정에서 겪는 어려움을 파악하여 주요 점검 결과와 보완 내용을 공개한다.

II. 용어 정의

본 윤리원칙에서 사용하는 주요 용어의 뜻은 다음과 같다.

- 인공지능(AI) : “인공지능(AI)”이란 학습, 추론, 지각, 판단, 언어의 이해 등 인간이 가진 지적 능력을 전자적 방법으로 구현한 것을 말한다.¹⁾
- AI 시스템 : “AI 시스템”이란 다양한 수준의 자율성과 적응성을 가지고 주어진 목표를 위하여 실제 및 가상환경에 영향을 미치는 예측, 추천, 결정 등의 결과물을 추론하는 AI 기반 시스템을 말한다.²⁾
- AI 전 생애주기 : “AI 전 생애주기”란 AI 시스템의 기획, 데이터 수집·구축·처리, 모델 개발·검증, 배포·제공, 운영, 모니터링·개선 및 종료에 이르는 전 과정을 말한다. 각 단계가 반드시 순차적으로 진행되는 것은 아니며, 특정 단계가 반복될 수도 있다.
- AI 행위자 : “AI 행위자”란 “AI 전 생애주기”에서 실질적인 역할을 하는 개인과 조직을 말한다.
- 이용자 : “이용자”란 AI 시스템 또는 AI가 적용된 제품·서비스를 제공받아 이용하는 개인과 조직을 말한다.
- 이해관계자 : “이해관계자”란 “AI 행위자”와 “이용자”를 포함하여 AI의 개발·제공·이용 등에 직접 또는 간접적으로 관여하거나 영향을 받는 모든 개인과 조직을 말한다.
- 위 개념은 서로 배타적이지 않다. 하나의 개인이나 조직은 사안과 역할에 따라 여러 지위에 함께 해당할 수 있다.

1) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제2조제1호

2) 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」 제2조제2호

Ⅲ. 가치

1. 인간의 존엄성

- 모든 사람은 그 존재 자체로 존중받으며, 그 존엄과 인권은 보장된다.
- 모든 사람은 자기 삶의 주인으로서 판단하고 선택하며, AI를 통해 그 기회를 넓힐 수 있다.
- 누구나 AI에 관한 공동의 결정에 주체적으로 참여할 수 있다.

인간의 존엄성은 AI를 어떻게 만들고 쓰든 인간이 그 중심에 있어야 한다는 지향점을 가리킵니다. 인간은 수단이 아닌 목적 그 자체로 대우받아야 합니다. 이는 한 사람의 삶이 다른 누군가의 목적에 따라 정해져서는 안 되며, 누구나 자기 삶의 주인으로 남아야 한다는 뜻입니다. 따라서 AI가 할 일은 사람의 판단을 대신하는 것이 아니라, 사람이 스스로 판단하고 선택하도록 뒷받침하는 것입니다. 물론 스스로 판단하고 선택할 범위가 개인의 일상에 국한되는 것은 아닙니다. 어떤 결정에 영향을 받는 사람이 그 결정 과정에 참여할 수 있어야 한다는 것도 자기에 관한 일을 스스로 정한다는 생각에서 나옵니다. AI를 어떻게 만들고 쓰고 평가할지 정하는 과정도 이 범위에 속합니다. AI의 영향을 받지 않는 사람이 거의 없는 오늘날, 이는 개인이 자기 일을 정하는 문제를 넘어 사회가 함께 결정해야 할 문제이기도 합니다. 그래서 AI에 관한 공동의 결정에는 누구나 주체적으로 참여할 수 있어야 하며, 이는 우리 사회의 민주적 의사 형성과도 이어집니다.

2. 사회의 공공선

- AI는 편익을 높이는 동시에 사회적 신뢰와 공동의 이익을 추구하는 방향으로 개발·활용되어야 한다.
- 우리 사회는 AI의 혜택을 더 많은 사람이 함께 누리는 공동체를 지향한다.
- AI의 개발과 활용은 다양한 구성원이 참여하는 공론장을 확대하고, 민주적 의사형성을 뒷받침하는 방향으로 이루어져야 한다.

사회적 공공선은 AI가 만들어 내는 이익과 그 방향에 관한 결정을 사회가 함께 나누는 상태를 지향합니다. AI의 개발과 활용은 사회적 신뢰를 해치지 않는 소극적 태도에 머무르지 않고, 편익과 공동의 이익을 함께 증진해야 합니다. 어떤 이익이 공동의 이익인지는 그 이익이 최종적으로 누구에게까지 미치는지에 좌우됩니다. AI의 혜택에 접근할 기회가 일부에 치우친다면 공동의 것이라 하기 어려우며, 더 많은 사람이 그 기회를 누릴 때 비로소 공동체라 부를 수 있습니다. 그러한 확산은 저절로 이루어지지 않습니다. 무엇을 함께 누릴 것인지를 우리 사회가 스스로 정할 때 비로소 가능해집니다. 여러 구성원이 참여하는 공론장이 확대될수록 사회는 스스로 판단할 역량을 갖추게 되고, 그렇게 모인 의견이 실제 결정으로 이어질 때 민주적 의사형성이 제 기능을 합니다.

3. 인류의 지속가능성

- AI는 미래 세대가 살아갈 환경을 보전하는 방향으로 개발·활용되어야 한다.
- 인류는 AI가 사회에 미치는 장기적 영향을 지속적으로 살피고, 그 변화에 책임 있게 대응해야 한다.
- AI는 민주주의를 강화하고, 국제 평화와 협력에 기여하는 방향으로 개발·활용되어야 한다.

인류의 지속가능성은 오늘의 AI가 남기는 환경적·사회적 부담을 다음 세대에 넘기지 않는 데 목표를 둡니다. 그 부담은 AI를 학습시키고 운영하는 데 드는 전력과 물에서 먼저 나타납니다. 그러므로 환경 보전은 여건이 허락할 때 하는 선택이 아니라 지금 세대가 이행해야 할 과제입니다. 부담이 환경에만 나타나는 것은 아닙니다. 일자리 구조처럼 한 세대 안에서는 잘 드러나지 않는 변화도 함께 살펴야 하며, 변화를 확인했다면 실제 대응이 뒤따라야 합니다. 그러한 대응은 어느 한 주체나 한 나라의 판단만으로 이루어질 수 없습니다. 부담이 여러 세대에 여러 나라에 걸쳐 있는 만큼, 이를 감당할 역량도 사회의 안과 밖에 함께 필요합니다. 안으로는 사회가 스스로 결정하는 민주주의가, 밖으로는 국가들이 갈등을 조정하는 평화와 협력이 그 바탕을 이룹니다.

IV. 원칙

① 인간중심성(Human-centeredness)

- AI 행위자는 AI가 사람의 판단과 선택, 창의성과 문제 해결 역량을 넓히는 방향으로 개발·활용되도록 노력한다.
- AI 행위자는 AI가 사람의 취약성을 부당하게 이용하거나 정보와 선택을 왜곡하여 자기결정을 침해하는 데 활용되지 않도록 한다.
- AI 행위자는 AI에 맡기는 판단의 범위와 사람의 역할을 정하고, AI의 활용 목적과 영향에 따라 사람이 그 작동을 파악하고 필요한 경우 개입할 수 있도록 지원한다.
- 이용자는 AI에 대한 과도한 의존을 경계하고, 그 산출물을 스스로 검토하여 주체적으로 활용하도록 노력한다.

AI는 사람의 지식을 따르는 도구를 넘어, 더 복잡한 판단과 업무에까지 활용되고 있습니다. 이러한 변화 속에서 사람과 AI가 어떤 관계를 맺어야 하는지를 다시 살펴볼 필요가 있습니다. 인간중심성은 AI가 사람의 삶을 더 나아지게 하고 사람의 능력과 가능성을 넓히는 방향으로 개발되고 활용되어야 한다는 원칙입니다. 사람은 AI를 활용해 미처 찾지 못한 대안을 얻거나 반복적인 업무의 부담을 줄이고 그만큼 더 중요한 판단과 창의적인 활동에 집중할 수 있습니다. 인간 중심성 원칙에서 중요한 것은 사람이 AI를 활용해 자신의 판단과 역량을 더 잘 발휘할 수 있는지입니다.

사람이 중심에 있으려면 스스로 판단하고 선택할 수 있어야 합니다. 사람의 불안이나 정보 부족과 같은 취약한 상태를 부당하게 이용하거나, 정보와 선택지를 왜곡해 특정한 결론으로 이끄는 데 AI가 활용되어서는 안 됩니다. 겉으로 선택할 수 있어 보여도 실질적으로 판단할 기회가 제한된다면, 사람이 중심에 있다고 보기 어렵습니다. AI가 맡는 업무의 범위가 넓어지고 작동 방식도 복잡해지면서 사람이 모든 과정을 감독하거나 판단하기는 어려워지고 있습니다. 그렇다면 사람의 역할은 개별 판단을 확인하는 것에서 AI가 일하는 방식 전반을 관리하는 것으로 옮겨져야 합니다. AI에 무엇을 어디까지 맡길지 미리 정하고 필요하면 그 작동을 제한하거나 중단할 수 있는 여지를 남겨 두어야 합니다. 이러한 수단은 AI를 설계하고 도입하는 단계에서 함께 고민할 필요가 있습니다. AI 행위자는 사람이 이러한 역할을 할 수 있도록 필요한 정보와 수단을 마련해야 합니다. 그 종류와 수준은 AI의 활용 목적과 영향에 따라 달라질 수 있습니다. 이용자의 역할도 중요합니다. 편리하다는 이유로 판단을 지나치게 AI에 맡기면, 스스로 생각하고 결정하는 힘이 약해질 수 있습니다. 이용자는 AI에 대한 과도한 의존을 경계하고, 그 산출물을 스스로 검토하여 자신의 판단과 목적에 맞게 활용하도록 노력해야 합니다.

[주체별 역할]

개발자 · 사업자	· AI가 사람의 판단과 선택을 지원하고, 창의성과 문제해결 역량을 넓히는 방향으로 설계·개발·운영한다.
	· AI가 사람의 취약성을 부당하게 이용하거나 정보와 선택지를 왜곡하여 특정한 판단이나 행동을 유도하지 않는지 점검한다.
사업자	· 사람이 AI에 맡긴 판단과 업무를 확인하고, 필요한 경우 제한·수정·중단하거나 그 결과에 대응할 수 있도록 필요한 정보와 기능·절차를 마련한다.
시민	· AI의 산출물을 자신의 상황과 목적에 맞게 검토하고, 중요한 사안에서는 다른 자료와 관점도 함께 살펴 스스로 판단한다.
	· AI에 판단을 지나치게 맡기지 않고, 자신의 판단과 역량을 넓히는 수단으로 주체적으로 활용한다.
	· AI를 다른 사람의 취약성을 이용하거나 정보와 선택을 왜곡하여 판단과 행동을 부당하게 유도하는 데 활용하지 않는다.
정부	· 국민이 AI의 기능과 한계를 비판적으로 이해하고 주체적으로 활용할 수 있도록 연령과 대상의 특성에 맞는 AI 문해력 교육과 정보를 제공한다.
	· 공공부문의 중요한 판단과 업무에 AI를 활용할 때, 사람이 그 과정과 결과를 확인하고 필요한 경우 개입하거나 대응할 수 있는 체계를 마련한다.
	· AI의 활용 목적과 영향에 따라 사람의 판단과 업무를 확인하고 개입하는 데 필요한 정보와 수단이 마련되도록 관련 기준과 지원 방안을 마련한다.
사회	· 언론·학계·시민사회 등은 AI가 사람의 판단과 선택, 창의성, 노동과 학습 등 인간의 역량에 미치는 영향을 지속적으로 살피고 검증한다.
	· AI가 사람의 취약성을 이용하거나 정보와 선택을 왜곡하는 사례를 비판적으로 살피고, 시민이 그 영향을 이해할 수 있도록 알기 쉬운 정보를 제공한다.
	· 사람이 AI를 주체적으로 활용하고 자신의 판단과 역량을 발전시키는 문화를 조성한다.

② 프라이버시 보호(Privacy)

- AI 행위자는 AI가 사생활을 침해하거나 부당한 감시에 활용되지 않도록 한다.
- AI 행위자는 AI 전 생애주기에서 개인정보를 정해진 목적에 필요한 범위에서 처리하고, 적절한 보호조치를 마련한다.
- AI 행위자는 정보주체의 개인정보자기결정권을 존중하고, 그 행사를 뒷받침한다.
- 이용자는 AI를 이용하는 과정에서 타인의 프라이버시를 존중한다.

개인정보의 유출을 막는 것만으로는 프라이버시를 충분히 보호할 수 없습니다. AI가 여러 정보를 결합·분석하는 과정에서, 중요하지 않아 보이던 정보로부터 개인의 성향, 생활방식, 건강 상태 등 새로운 정보가 추론될 수 있기 때문입니다. 이러한 정보가 개인이 예상하지 못한 방식으로 활용되면, 개인은 부당한 감시나 조작의 대상이 될 수 있고 자신의 정보가 어떻게 쓰이는지 통제하기도 어려워집니다.

프라이버시를 보호하려면 개인정보를 정해진 목적에 필요한 범위에서만 처리해야 합니다. 개인정보의 유출과 오용뿐 아니라 불필요한 식별과 민감정보의 노출도 막아야 하며, 정보주체가 자신의 개인정보가 어떻게 처리되고 있는지 분명히 알고, 정정이나 삭제 등을 요구할 수 있어야 합니다.

이용자가 업무나 일상에서 타인의 개인정보를 입력하는 경우에도 그 정보가 학습이나 다른 목적에 활용될 수 있습니다. 이용자는 자신이 다루는 정보가 누구의 것인지, 그 정보를 AI에 입력해도 되는지 살펴봐야 합니다.

[주체별 역할]

개발자 · 사업자	· 데이터의 출처와 이용 권한, 수집 목적과 범위, 보유 기간을 점검한다. · 보유 데이터의 가명·비식별 처리, 접근 통제 등 필요한 보호조치를 갖춘다. · 민감정보나 사적인 정보가 학습데이터 또는 AI 산출물을 통해 불필요하게 노출되지 않도록 점검한다. · 개인정보 침해가 발생한 경우 신속히 대응·시정할 수 있는 절차를 마련한다.
시민	· 다른 사람의 개인정보, 사적인 대화, 사진, 음성, 위치 정보와 민감한 정보 등을 당사자의 동의 없이 AI에 입력하거나 공유하지 않는다. · 자신의 개인정보와 창작물 등이 AI에 어떻게 수집·이용되는지 살핀다. · 자신의 개인정보에 대해 필요한 경우 동의 철회, 삭제, 정정 등 보장된 권리를 행사한다.
정부	· AI가 부당한 감시나 사생활 침해의 수단으로 남용되지 않도록 법·제도를 정비한다. · 개인정보 침해 시 신고·조사·시정과 피해구제가 실질적으로 이루어지도록 지원한다. · 공공 부문에서 AI를 활용하는 경우 개인정보의 필요성과 처리 범위를 엄격히 살피고, 국민의 정보를 안전하게 보호할 수 있는 관리 체계를 마련한다.
사회	· 언론·학계·시민단체 등은 AI를 통한 감시, 개인정보 오남용과 사생활 침해 실태를 독립적으로 조사·감시하고, 사회적으로 필요한 개선 사항을 제안한다. · 개인정보를 보호하면서도 공익을 위한 연구·보도가 가능하도록 관련 기준과 사회적 논의를 발전시키도록 노력한다.

③ 공정성·포용성(Fairness and Inclusiveness)

- AI 행위자는 AI의 편향이 개인이나 집단에 대한 차별이나 부당한 불이익으로 이어지지 않도록 한다.
- AI 행위자는 사회적 약자를 포함한 누구나 AI에 접근하고, 그 혜택과 기회를 공평하게 누릴 수 있도록 노력한다.
- AI 행위자는 AI의 개발·활용·평가에 관한 논의에서 다양한 이해관계자의 주체적 참여를 존중하고, 이를 뒷받침하도록(보장하도록) 노력한다.
- 이용자는 AI를 차별과 혐오를 확산하는 데 활용하지 않으며, AI 산출물에 담긴 편견을 그대로 받아들이거나 퍼뜨리지 않는다.

AI는 학습에 사용한 데이터나 설계 방식에 따라 특정 개인이나 집단에 편향된 결과를 낼 수 있습니다. 기존 사회의 차별과 불평등이 데이터에 반영되어 있다면, AI의 판단이나 산출물이 이를 되풀이하거나 확대할 수도 있습니다. 특정 집단의 특성이나 경험이 데이터에 충분히 담기지 않은 경우에도 그 집단에 대해서는 결과의 정확성과 품질이 낮아질 수 있습니다. 또한 필요한 기기나 이용 환경을 갖추기 어렵거나 AI를 활용할 역량이 부족한 사람은 AI에 접근하거나 그 혜택과 기회를 누리는 데 제약을 받을 수 있습니다.

편향을 완전히 없애기는 어렵고, 무엇이 부당한 차별인지는 AI가 활용되는 맥락에 따라 달라질 수 있습니다. 따라서 데이터와 알고리즘, 산출물에 나타나는 편향을 지속적으로 점검하고 개선해야 합니다. 이때 특정 집단이 데이터에 충분히 반영되었는지도 함께 살펴볼 필요가 있습니다. 공정성·포용성 원칙은 편향의 존재 자체를 완전히 제거할 수는 없더라도 그것이 차별이나 부당한 불이익으로 이어지지 않도록 하는 데 초점을 둡니다.

누구나 AI의 혜택과 기회를 공평하게 누릴 수 있어야 한다는 것은 모두에게 같은 결과를 보장한다는 뜻이 아닙니다. 기기나 이용 환경, 경제적 여건, 디지털 역량 또는 AI가 설계된 방식 등으로 인해 AI에 접근하거나 그 혜택과 기회를 누리는 과정에서 부당하게 배제되지 않아야 한다는 의미입니다. 특히 장애인, 고령자, 저소득층 등 AI 제품이나 서비스를 이용하는 데 어려움을 겪는 사람에게 실제 이용에 필요한 지원이 제공되어야 합니다.

AI가 사회에 미치는 영향이 커질수록, AI의 개발·활용·평가에 관한 논의에는 다양한 관점과 경험이 반영되어야 합니다. AI의 영향을 크게 받거나 그동안 논의에서 소외되기 쉬웠던 집단의 관점이 반영될 때 편향과 배제의 문제를 미리 발견할 수 있습니다. 이는 다양한 이해관계자가 개별 기업의 개발 과정에 직접 참여해야 한다는 뜻이 아닙니다. 시민을 비롯한 이해관계자가 AI가 어떻게 활용되고 사회에 어떤 영향을 미치는지에 관한 논의에 주체적으로 참여할 수 있어야 한다는 의미입니다.

[주체별 역할]

개발자 · 사업자	· 데이터의 수집·구축·가공·라벨링 단계부터 인공지능의 결과까지 특정 개인이나 집단에 불리한 편향 가능성을 살핀다.
	· 서비스 운영 과정에서 차별적이거나 불리한 결과가 반복되는지 점검·개선한다.
시민	· 장애, 연령, 언어, 지역과 이용 환경 등 다양한 특성을 고려하여 누구나 쉽게 AI를 이용할 수 있도록 설계·운영한다.
	· AI의 영향을 받는 당사자와 이해관계자의 의견이 설계·운영·개선 과정에 반영될 수 있도록 노력한다.
정부	· AI를 활용하여 특정 개인이나 집단을 차별·배제하거나 혐오와 편견을 확산하지 않는다.
	· AI의 결과에 담긴 편견이나 차별을 그대로 믿거나 퍼뜨리지 않는다.
사회	· AI의 개발과 활용으로부터 발생하는 차별이나 불이익을 발견하면 알리고, 공론과 의사결정에 반영될 수 있도록 참여한다.
	· 공공, 노동, 복지, 교육, 금융, 치안 등 국민의 삶에 큰 영향을 미치는 영역에서 AI로 인한 차별과 불이익이 발생하지 않도록 점검한다.
정부	· 영향을 받는 당사자가 의견을 제시하고 참여할 수 있는 절차를 마련한다.
	· 지역·소득·장애·연령과 이용 역량 등에 따른 격차를 줄이기 위한 지원을 제공한다.
사회	· 사회적 약자를 지원하는 공공서비스와 돌봄·복지 영역에서 AI가 적절히 활용될 수 있도록 지원 체계를 마련한다.
	· 언론·학계·시민단체 등은 AI가 사회적 차별과 불평등을 새롭게 만들거나 확대하는지 독립적으로 조사·검증하고, 편향과 차별을 개선하기 위한 대안을 제시한다.
사회	· 사회적 논의에 소수자와 사회적 약자를 포함한 다양한 시민의 경험과 관점이 드러나도록 공론과 소통의 기회를 마련한다.

④ 책임성(Accountability)

- AI와 관계되는 모든 주체는 AI와 관련된 책임이 불분명해질 수 있음을 인식하고, 각자의 역할과 영향력에 맞는 책임을 다하도록 노력한다.
- AI 행위자는 문제가 발생한 경우 시정과 피해구제가 실질적으로 이루어지도록 한다. 또한 책임 소재를 밝히기 위해 노력한다.
- AI 역량과 자원을 많이 보유한 주체는 그에 상응하는 책임을 지며, 국제사회에서 AI의 혜택이 특정 국가·지역·집단에 집중되지 않도록 노력한다.

AI의 개발과 활용에는 여러 주체가 관여합니다. 데이터를 모으고 모델을 만드는 주체, 이를 제품이나 서비스로 제공하는 주체, 이를 운영하고 이용하는 주체가 각각 다릅니다. 그래서 문제가 발생했을 때 누구에게 책임이 있는지 곧바로 가려지지 않을 수 있고, 그사이 피해를 입은 사람이 필요한 도움을 받지 못할 수 있습니다.

책임이 불분명해질 수 있다는 사실이 책임을 미룰 이유가 되지는 않습니다. AI의 영향이 사회 전반에 미치는 만큼, 그 책임도 어느 한 주체에겐 돌아가지 않습니다. AI를 만들고 제공하는 주체, 이를 활용하는 주체, 관련 제도와 여건을 마련하는 주체 등이 각자가 맡은 역할과 영향력에 맞는 책임을 다해야 합니다. 이는 모든 이해관계자에게 같은 수준의 책임을 요구한다는 뜻이 아니라, 각자의 역할과 영향력에 따라 책임의 범위가 달라진다는 의미입니다.

문제가 발생한 경우에는 시정과 피해구제가 실질적으로 이루어져야 합니다. 책임 소재를 밝히는 일도 중요하지만, 책임 소재가 밝혀질 때까지 피해구제가 미뤄져서는 안 됩니다. 여러 주체가 얹혀 있어 책임을 규명하는 데 시간이 걸릴 수 있는 만큼, 이 원칙은 피해구제가 책임 규명의 결과에 좌우되지 않도록 하는 데 의미가 있습니다.

AI 역량과 자원을 많이 보유한 주체는 그만큼 큰 영향력을 행사할 수 있으므로 그에 상응하는 책임을 져야 합니다. 이 책임은 국내에 한정되지 않습니다. AI의 혜택이 일부 국가나 지역, 집단에만 돌아간다면 기존의 격차가 오히려 더 벌어질 수 있기 때문입니다.

[주체별 역할]

개발자 · 사업자	· AI의 기획·개발·제공·운영 과정에서 각 담당자의 권한과 책임을 명확히 하고, 중요한 결정과 변경 사항을 기록·관리한다.
	· 문제가 발생하면 신속히 바로잡고 피해구제에 협조할 수 있는 내부 절차를 마련한다. · AI 역량과 데이터·자원, 시장 영향력을 많이 보유한 사업자는 그에 상응하는 책임을 인식하고 행동한다.
시민	· AI를 이용할 때 자신의 행동이 다른 사람과 사회에 미칠 영향을 생각한다. · AI를 허위정보 유포, 저작권 침해, 명예훼손, 차별과 사생활 침해 등에 이용하지 않는다. · AI로 인해 자신의 권리나 이익이 침해되면 설명과 재검토를 요청하고 피해구제 절차를 활용한다. · 문제나 개선이 필요한 사항을 발견하면 관련 기관이나 서비스 제공자에게 알린다.
정부	· AI 관련 사고나 권리침해가 발생했을 때 누가 어떤 책임을 져야 하는지 확인할 수 있는 제도 마련한다. · 이의제기·재검토·분쟁조정·피해구제 절차가 실질적으로 작동하도록 법·제도와 지원체계를 마련한다. · 공공기관의 AI 도입·운영 과정에서도 의사결정과 책임 구조를 명확히 하고, 조달이나 위탁을 이유로 공공의 책임이 민간에 전가되지 않도록 관리한다. · 국가 간 AI 격차를 줄이고 혜택을 넓게 나누기 위한 국제협력을 추진한다.
사회	· 언론·학계·시민단체 등은 기업과 정부가 AI를 책임 있게 개발·활용하는지 독립적으로 모니터링한다. · AI와 관련한 문제가 발생하면 원인과 책임이 충분히 논의될 수 있도록 충분한 근거와 함께 사회에 알린다. · 다양한 시민과 이해관계자가 AI 관련 정책과 제도에 참여할 수 있는 기회를 마련하고, 책임 있는 AI 활용을 위한 사회적 기준과 자율규범의 형성을 지원한다.

⑤ 안전성(Safety)

- AI 행위자는 AI로 인해 사람의 생명·신체·건강에 위해가 발생하거나 재산상 피해가 생기지 않도록 한다.
- AI 행위자는 각자의 역할에 따라 AI의 위험 수준에 맞는 안전조치를 마련하고 오남용이나 외부 공격 등에 대비한다. 피해가 발생한 경우에는 지체 없이 그 확대를 방지한다.
- 이용자는 AI를 타인의 생명·신체·건강·재산에 위해를 가하는 데 활용하지 않는다.

자율주행차가 사람이나 장애물을 잘못 인식하면 사고가 발생할 수 있습니다. 의료 현장에서 AI의 판단 결과에 오류가 있으면 환자의 건강이 위협받을 수 있고, 산업설비를 제어하는 AI가 비정상적으로 작동하면 작업자가 다치거나 시설이 파손될 수 있습니다. 또한 해킹이나 무단 접근으로 AI의 작동이 변경되면 사람과 재산에 피해가 발생할 수 있습니다. AI의 안전성은 이러한 위험으로부터 사람의 생명·신체·건강과 재산을 보호하는 문제입니다. 여기서 재산상 피해는 AI의 작동이나 오남용, 외부 공격으로 발생하는 직접적인 피해를 말하며, 저작권 등 권리 침해는 관련 원칙과 법제에서 다룹니다.

모든 AI에 같은 수준의 안전조치가 필요한 것은 아닙니다. AI의 용도와 활용 환경, 발생할 수 있는 피해의 성격과 정도에 따라 필요한 조치가 달라집니다. 또한 위험은 AI 자체의 오류나 이상 작동에서만 발생하는 것이 아닙니다. AI가 본래의 목적과 다르게 사용되거나 외부의 공격을 받는 경우에도 피해가 발생할 수 있습니다. 모든 사고를 완전히 막기는 어려운 만큼, 피해가 발생한 경우에는 확대되지 않도록 신속히 대응해야 합니다.

오남용은 AI를 만드는 쪽의 대비만으로 막을 수 있는 것이 아닙니다. 안전을 위해 마련된 장치를 우회하려 하거나, AI를 다른 사람에게 해를 끼치는 목적에 사용하지 않는 것은 이용자가 스스로 지켜야 할 영역입니다. 특히 AI를 이용해 타인의 생명이나 신체, 재산에 해를 가하는 행위는 어떠한 경우에도 허용되지 않습니다.

[주체별 역할]

개발자	· AI의 용도와 활용 환경, 발생 가능한 피해의 성격과 정도를 고려하여 위험 수준에 맞는 안전조치를 마련한다.
사업자	· 시스템 오류와 외부 공격, 예상 가능한 오용과 악용에 대비한 안전·보안 조치를 마련한다. · 이상 징후나 사고가 발생하면 AI의 작동을 제한·중단하거나 수정할 수 있도록 대응·복구 절차를 마련한다.
시민	· 자신의 계정과 접근 권한을 안전하게 관리하고 AI 시스템의 보안·안전장치를 우회하거나 침해하지 않는다. · AI를 다른 사람과 사회의 안전을 해칠 수 있는 방식으로 사용하지 않는다. · AI의 오류나 이상한 작동, 위험한 결과 또는 악용 사례를 발견하면 관련 사업자나 기관에 알린다. · 안전성이 확인되지 않은 결과물을 무분별하게 이용·확산하지 않는다.
정부	· 분야별 특성과 위험 수준에 맞는 AI 안전 기준, 시험·평가와 사고 대응체계를 마련한다. · 중소기업과 공공기관 등이 필요한 안전조치를 갖추 수 있도록 기술·인력·정보를 지원한다. · AI의 새로운 위험과 사회적 영향을 지속적으로 조사하고, 필요한 제도를 개선한다. · 중대한 사고에 신속히 대응할 수 있도록 관계기관 간 정보공유와 협력 체계를 갖춘다.
사회	· 언론·학계·연구기관과 시민단체 등은 AI의 안전성과 사회적 영향을 독립적으로 연구·검증한다. · AI 사고와 위험에 관한 정보를 정확하고 책임 있게 전달한다. · 확인되지 않은 위험을 지나치게 과장하거나 축소하지 않으면서, 새로운 위험과 대응방안을 함께 논의할 수 있는 사회적 기반을 마련한다.

⑥ 신뢰성(Reliability)

- AI 행위자는 AI의 이용 목적과 적용 범위를 고려하여 필요한 수준의 성능을 정하고, 이를 갖추도록 노력한다.
- AI 행위자는 AI가 의도된 목적과 허용된 권한의 범위 안에서 작동하도록 설계한다.
- AI 행위자는 AI가 다양한 사용 조건에서 안정적으로 작동하도록 노력하고, 이용되는 동안 그 성능과 작동 상태를 지속적으로 점검한다.

AI에 요구되는 성능은 절대적으로 정해지는 것이 아닙니다. 같은 정도의 오차라도 AI의 이용 목적과 적용 분야, 그 결과가 미치는 영향에 따라 허용될 수 있고 그렇지 않을 수도 있습니다. 따라서 AI의 사용 목적과 적용 범위를 고려하여 어느 정도의 성능이 필요한지를 먼저 정하고, 이를 갖추도록 노력해야 합니다.

AI가 맡은 업무를 신뢰할 수 있게 수행하려면 의도된 목적과 허용된 권한의 범위가 분명해야 합니다. AI가 그 범위를 벗어나 작동하거나 비정상적으로 다른 시스템에 접근하면 의도한 기능과 다른 결과가 나타날 수 있습니다. 이는 AI의 모든 행동과 결과를 미리 예측해야 한다는 뜻은 아니며, AI가 맡은 업무와 부여받은 권한의 범위 안에서 작동하도록 설계해야 한다는 의미입니다.

같은 기능을 수행하도록 설계된 AI라도 입력되는 데이터나 이용 환경이 달라지면 성능과 결과의 품질이 낮아질 수 있습니다. 시험 환경에서는 안정적으로 작동하던 AI가 실제 현장에서는 필요한 성능을 내지 못할 수 있고, 이용되는 동안 데이터나 이용 방식이 달라지면서 성능이 서서히 저하될 수도 있습니다. 여기서 안정적으로 작동한다는 것은 AI가 언제나 동일한 결과를 내야 한다는 뜻은 아닙니다. 생성형 AI처럼 결과가 달라지는 것이 본래의 특성인 경우도 있으므로, 결과의 동일성보다 의도된 목적을 수행하는 데 필요한 기능과 성능이 다양한 조건에서도 유지되는지가 중요합니다.

AI의 성능과 작동 상태는 이용 전에 한 번 확인하는 것만으로는 충분하지 않습니다. 실제 사용 환경에서 발생할 수 있는 모든 상황을 미리 예상하기는 어렵고, 시간이 지나면서 데이터와 이용 환경도 달라질 수 있기 때문입니다. 따라서 AI가 이용되는 동안에도 그 성능과 작동 상태를 지속적으로 점검해야 합니다. 이러한 변화를 제때 발견해야 이용자는 AI의 결과를 어느 정도까지 신뢰하고 활용할 수 있는지 판단할 수 있습니다.

[주체별 역할]

개발자 · 사업자	· AI의 사용 목적과 적용 범위를 고려하여 필요한 수준의 성능을 정하고, 이를 설계·개발에 반영한다. · AI가 의도된 목적과 허용된 권한의 범위를 벗어나지 않도록 접근 권한과 작동 범위를 설정·점검한다. · 다양한 사용 조건에서 AI의 성능과 작동 상태를 확인하고, 이용되는 동안 성능 저하나 이상 작동을 지속적으로 살핀다.
시민	· AI가 언제나 동일하거나 정확한 결과를 제공하는 것은 아님을 이해한다. · 안내된 이용 목적과 권장 범위를 벗어난 방식으로 무리하게 사용하지 않는다. · 평소와 다른 결과나 예상하지 못한 작동이 반복되는 경우 이를 그대로 신뢰하거나 확산하지 않고, 관련 사업자나 기관에 알린다.
정부	· AI의 성능과 작동 범위를 평가할 수 있는 분야별 기술표준과 품질평가 기반을 마련한다. · 공공부문에서 활용하는 AI의 성능과 한계를 정기적으로 점검한다. · 국민에게 중대한 영향을 미치는 AI 시스템이 변경되거나 성능이 떨어질 경우 대응할 수 있는 절차를 마련한다.
사회	· 언론·학계·연구기관과 시민단체 등은 AI의 실제 성능과 한계를 독립적으로 연구·검증한다. · 예상하지 못한 결과나 사회적 영향이 확인되면 그 결과를 사회에 충분한 근거와 함께 알린다. · AI에 대한 과도한 기대나 불안을 경계하고, 기술의 실제 작동과 근거를 바탕으로 사회적 이해와 논의가 이루어지도록 노력한다.

⑦ 투명성(Transparency)

- AI 행위자는 각자의 책임을 다하는 데 필요한 정보를 공유하고, 이용자와 영향을 받는 사람에게 AI 활용 사실과 역할·한계 등을 명확하고 알기 쉽게 제공한다.
- 이용자는 제공된 정보를 확인하고 자신이 이용하는 제품이나 서비스에서 AI가 어떤 역할을 하는지 이해하기 위해 노력한다.
- AI 행위자는 생명·신체의 안전과 기본권을 포함한 개인의 권리·의무·기회에 중대한 영향을 미치는 결정에 AI가 활용되는 경우, 당사자가 그 결정을 이해하고 필요한 대응을 할 수 있는 여건을 마련하기 위해 노력한다.

AI가 활용된다는 사실을 알리는 것만으로는 투명성이 충분히 확보되지 않습니다. 이용자와 영향을 받는 사람은 AI가 어떤 목적으로 활용되는지, 어떤 역할을 하며 어떤 한계가 있는지 이해할 수 있어야 합니다. 이를 위해서는 AI 행위자 사이에서도 필요한 정보가 공유되어야 합니다. AI를 개발한 주체와 이를 제품이나 서비스로 제공하는 주체, 실제로 운영하는 주체가 서로 다른 경우가 많기 때문입니다. 각 주체가 자신의 역할과 책임을 다하려면 AI의 기능과 한계, 작동 상태 등에 관한 필요한 정보를 서로 공유해야 합니다.

투명성을 확보하려면 정보를 제공하는 주체의 노력뿐 아니라 이용자의 능동적인 확인도 필요합니다. AI가 눈에 잘 띄지 않는 방식으로 활용되는 경우가 있으므로, 이용자도 제공된 정보를 확인하고 자신이 이용하는 제품이나 서비스에 AI가 어떻게 활용되는지 스스로 확인하려는 노력이 필요합니다. 이용자의 이러한 노력이 정보를 제공할 책임을 대신하지는 않습니다.

대출 심사나 채용 평가처럼 개인의 권리·의무·기회에 중대한 영향을 미치는 결정에 AI가 활용되는 경우에는 당사자를 위한 별도의 보호가 필요합니다. 여기서 말하는 중대한 영향은 법령이 정하는 고영향 AI의 범위에 한정되지 않습니다. 개인의 권리와 기회에 영향을 미치는 AI의 활용은 계속 넓어지고 있으므로, 원칙은 그보다 넓은 범위를 포괄합니다. 특히 사람의 개입 없이 AI만으로 결정이 이루어지는 경우에는 더욱 그렇습니다. 여기서 필요한 대응에는 해당 결정에 관한 설명을 요구하는 것, 그 결정에 이의를 제기하는 것, 사람에 의한 재검토를 요청하는 것이 포함됩니다.

AI의 학습 과정도 투명성과 밀접하게 연결됩니다. 저작물이나 그 밖의 보호받는 자료가 AI의 학습에 활용되는 경우에는 학습 데이터의 출처와 활용 방식, 관리관계에 관한 필요한 정보를 공개하는 것이 중요합니다. 저작자와 관리자가 자신의 저작물이 AI 학습에 어떻게 이용되었는지 확인할 수 있어야 권리침해에 대응할 수 있고, 지식재산권도 실질적으로 보호될 수 있습니다. 모든 정보를 같은 수준으로 공개해야 한다는 뜻은 아닙니다. 제공하거나 공개할 정보의 범위와 구체성은 AI의 활용 목적과 영향, 정보를 제공받는 사람의 특성 등에 따라 달라질 수 있습니다. 개인정보와 영업비밀 등 보호할 필요가 있는 정보도 함께 고려해야 합니다. 그러나 이러한 사정이 중대한 영향을 받는 당사자에게 필요한 설명을 일률적으로 거부하는 이유가 되어서는 안 됩니다.

[주체별 역할]

개발자 · 사업자	· 이용자가 AI와 상호작용하거나 AI가 생성한 결과물을 접하는 경우 AI 활용 사실을 알기 쉽게 표시한다.
	· AI의 역할·기능·한계와 유의사항을 이해하기 쉬운 방식으로 제공한다.
시민	· 개인의 권리나 이익에 중대한 영향을 미치는 결정에 AI가 활용되는 경우, 당사자가 그 의미와 주요 근거를 이해할 수 있도록 설명하고 이의제기 절차를 마련한다.
	· 데이터와 정보를 공개할 때에는 개인정보와 보안, 영업비밀, 지식재산권도 함께 보호한다.
정부	· 자신에게 중대한 영향을 미치는 결정에 AI가 활용된 경우 필요한 설명과 관련 정보의 확인을 요청한다.
	· AI를 활용한 결과에 문제가 있다고 생각되면 이의제기나 재검토 절차를 활용한다.
사회	· AI를 이용하여 콘텐츠나 정보를 생산·공유할 때 다른 사람이 AI 활용 여부를 오인할 가능성이 있는 경우 인공지능을 활용한 사실을 알린다.
	· AI가 생성한 내용을 확인 없이 사실로 단정하거나 퍼뜨리지 않는다.
정부	· 분야별 특수성, 위험 수준과 국민에게 미치는 영향, 개인정보·보안·지식재산권 보호 필요성을 고려하여 투명성의 세부 기준과 절차를 마련한다.
	· 공공부문에서 AI를 활용하는 경우 국민이 AI의 활용 목적과 범위, 담당 기관과 이의제기 방법을 알 수 있도록 공개한다.
사회	· 국민이 설명을 듣고 정보를 확인하며 이의를 제기할 수 있는 절차가 형식적으로 운영되지 않도록 관리한다.
	· 언론·학계·시민단체 등은 기업과 정부가 AI 관련 정보를 정확하고 충분하게 공개하고 있는지 모니터링 한다.
사회	· 공개된 정보가 사실에 맞고 시민이 이해하기 쉬운지 검증한다.
	· 시민이 AI의 기능과 한계, 사회적 영향을 이해할 수 있도록 정확한 정보와 분석을 제공하고, AI 생성물과 허위·조작정보를 책임 있게 검증하는 환경을 조성한다.