

인공지능 책임성과 공공성 강화를 위한 시민사회 공동행동

수 신 각 언론사 정치부·사회부

발 신 인공지능 책임성과 공공성 강화를 위한 시민사회 공동행동
(담당: 디지털정의네트워크 희우 활동가 02-774-4551, 참여연대 공익법센터 이지은 간사 02-723-0666)

제 목 [보도자료] AI시민행동, 인공지능 윤리원칙(안)에 대한 의견서 제출

날 짜 2026. 7. 9. (총 7 쪽)

보 도 자 료

AI시민행동, 인공지능 윤리원칙(안)에 대한 의견서 제출

국제 기준, 2020년 윤리기준보다도 후퇴, 인권 보호와 피해 구제 대책
강화해야

영향받는 사람의 피해 위험 방지하고 구제 가능성 높여야
'인권기반 접근' 없는 AI 윤리원칙(안) 전면 개선 촉구

1. 어제(7/8) 인권, 노동, 복지, 여성, 환경, 소비자, 평화 분야 전국 48개 시민사회단체들이 함께 하고 있는 <인공지능 책임성과 공공성 강화를 위한 시민사회 공동행동(이하 'AI시민행동')>은 과학기술정보통신부와 정보통신정책연구원이 공개한「[인공지능 윤리원칙\(안\)](#)」에 대한 의견서를 제출했다.
2. 이번 인공지능 윤리원칙(안)은 [2020년 제정된「인공지능 윤리기준](#)」을 대체하는 것으로, 관계부처 협의와 사회 각계 의견 수렴을 거쳐 국가AI전략위원회에서 최종 확정될 예정이다. AI 시민행동은「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 인공지능기본법)이 이미 시행 중인 상황에서, 이번 윤리원칙이 국제적 정합성을 높이고 사업자의 준수 기준을 제시하려면 무엇보다 영향받는 사람의 피해 위험을 방지하고 구제 가능성을 높이는 방향이어야 한다고 강조했다. 또한 인공지능 기술로 인해 우리 사회 대전환이 예상되고 그에 따른 윤리원칙을 담고자 한다면 윤리원칙의 철학과 방향

그리고 인권, 성평등, 공공성 등 인류 보편의 가치와 규범을 더욱 명확하게 규정할 필요가 있다고 지적했다.

3. 첫째, 인공지능에 대한 인권기반 접근을 윤리원칙에 명시할 것을 요구했다. 유엔 인권기구는 2018년 이후 일관되게 [‘인공지능에 대한 인권기반 접근’](#)을 국제기준으로 제시해 왔고, 2024년 채택된 [유럽평의회 <인공지능과 인권, 민주주의 및 법치주의에 관한 기본 협약>](#)도 인권 준수를 명시적으로 요구하고 있다. 반면 이번 윤리원칙(안)의 3대 가치와 6대 원칙 어디에도 ‘인권’이라는 표현이 등장하지 않으며, ‘인간의 존엄성’ 가치가 침해 대상을 ‘생명·신체·건강·재산’으로 열거한 것은 물리적·경제적 위해에만 초점을 맞추고 있다. 그 결과 인공지능을 매개로 한 젠더 기반 폭력과 차별이 초래하는 정신적·사회적 위해를 포착하지 못한다. ‘인간의 존엄성’ 가치 항목에 [세계인권선언](#), [여성차별철폐협약\(CEDAW\)](#) 등 국제인권법이 인정하는 인권의 보호·존중과 인권침해를 구제할 기업의 책임을 명시해야 한다고 요구했다.
4. 둘째, 인공지능기본법이 다루지 못하는 위험을 보완할 ‘윤리적 금지선(Ethical Red Line)’을 윤리원칙에 명문화할 것을 요구했다. 현행 인공지능기본법은 금지하는 인공지능을 별도로 규정하고 있지 않아 국제적으로 금지된 완전자율살상무기의 개발·이용이나 위험·위법적으로 직접적·간접적 차별을 낳는 인공지능을 규율하지 못하고 있다. 아울러 공공장소에서의 실시간 식별·추적, 감정인식 등 인권적으로 용인할 수 없는 인공지능 역시 금지하고 있지 않다. AI시민행동은 윤리원칙의 사회의 공공성 가치 항목에 국제 인권기준상 금지되거나 인권과 양립할 수 없는 위험을 야기하는 인공지능을 예시적으로 열거하고, 그 개발과 이용을 금지하도록 권고할 것을 요구했다.
5. 셋째, 영향받는 사람의 알 권리와 참여권을 보장하고 차별을 적극적으로 시정할 것을 요구했다. 고용, 교육, 보건의료, 사회복지 등 삶과 노동의 중대한 영역에서 인공지능에 기반한 결정이 확산되면서 안전과 인권에 중대한 영향을 미치거나 인권침해적·차별적 결과를 낳을 수 있다는 것이다. 또한 영향받는 사람 보호와 관련한 법률적 공백을 지적한 [감사원 실지감사\(1/5\)](#)를 언급하며, 인공지능 사업자가 영향받는 사람의 인권침해를 구제하는 책임성을 갖춰야 한다고 강조했다. 특히 윤리원칙(안)의 투명성 원칙이 “프라이버시, 영업비밀 보호 등 다른 정당한 가치와의 균형 속에서 구현되어야 한다”고 단서를 단 것에 대해 영업비밀 보호를 투명성 제한 사유로 명시하는 것은 매우 부적절하며 투명성이 기업의 영업비밀 보호 논리에 밀려 형해화될 우려가 있다고 비판했다.
6. 넷째, ‘공정성·포용성’ 원칙을 보다 구체화하고 ‘공공성’ 원칙을 되살릴 것을 요구했다. 다양성, 공정성, 포용성의 의미를 너무 협소한 관점으로 서술하고 있어 각각의 정의를 차별금지와 평등의 관점에서 서술해야 할 것을 강조했다. 현재 윤리원칙(안)은 데이터와 알고리즘의 편향에 대해 “지속적인 점검과 개선이 이루어져야 한다”는 추상적이고 소극적인 권고에 그치고 있다. 성별, 장애, 인종, 연령 등에 대한 역사적·구조적 차별이

인공지능 시스템을 통해 재생산·고착화되는 것을 막기 위해 '비차별(차별금지)'을 원칙에 명확히 명시할 것을 요구하였다. 또한 직접적·간접적 차별 결과가 사람에게 영향을 미친 경우 인공지능사업자가 그 시정과 구제에 나서는 것이 마땅한 윤리적 책임이며, 이를 윤리원칙에 명시해야 한다고 밝혔다. 더불어 인공지능은 공공의 이익과 사회적 가치를 증진하고, 민주주의와 인권을 강화하는 방향으로 개발 및 활용되어야 하며, 사회 전체의 이익과 공동체의 지속가능한 발전에 기여하도록 2020년 윤리기준에서 명시된 '공공성' 원칙도 되살려야 한다고 밝혔다.

7. 다섯째, 인공지능사업자의 '책임성' 원칙을 되살릴 것을 요구했다. 2020년 윤리기준과 OECD, EU 등 국제기준은 모두 책임성을 핵심 요건으로 두고 있으나, 이번 윤리원칙(안)은 EU의 7대 요건 중 책임성이 빠져 있다. 정부가 발표한 [<대한민국 인공지능 행동계획>](#)도 윤리원칙 고도화의 배경으로 "AI 기술이 일상과 산업 전반에 깊숙이 통합되면서 알고리즘의 책임성, 투명성, 공정성 등 사회적 신뢰를 확보하기 위한 윤리 체계의 고도화가 필수 과제로 부상하고 있다"고 언급한 바 있다. 책임의 귀속을 다루는 원칙이 없다는 것은 피해 발생 시 책임 소재와 구제 근거를 윤리원칙에 담지 않겠다는 것과 다름없다고 비판하며, 편향으로 인한 부당한 차별·배제와 피해 구제 및 보호를 위해서 책임성 원칙 부활과 독립적 점검·감독 체계 마련을 촉구했다.
8. AI시민행동은 "인공지능 윤리원칙이 국제적 정합성을 갖추고 사업자의 바람직한 준수 기준으로 기능하려면 무엇보다 영향받는 사람의 피해 위험을 방지하고 구제 가능성을 높이는 방향이어야 한다"며 정부가 해당 의견서 내용을 윤리원칙(안)에 반영할 것을 촉구했다. 끝.

▣ 붙임 1. 인공지능 윤리원칙(안)에 대한 의견서 전문 (총 5쪽)

인공지능 윤리원칙(안)에 대한 의견서

인공지능 윤리원칙은 국제기준의 충실한 반영으로
영향 받는 사람의 피해 위험 방지하고 구제 가능성 높여야
인공지능사업자의 '책임성' 삭제 등 일부 규정은 2020년 윤리기준보다도 후퇴

과학기술정보통신부와 정보통신정책연구원은 2020년 제정된「인공지능 윤리기준」을 잇는 새로운「인공지능 윤리원칙(안)」을 공개하고 의견을 수렴하고 있습니다. 인공지능 윤리원칙(안)은 관련 부처 협의와 사회 각계 의견 수렴 절차를 거쳐 국가AI전략위원회에서 최종 확정될 예정입니다. 이에 인권, 노동, 복지, 여성, 환경, 소비자, 평화 분야 전국 46개 시민사회 단체들이 함께 구성한 <인공지능 책임성과 공공성 강화를 위한 시민사회 공동행동(이하 'AI시민행동')>은 인공지능 윤리원칙(안)에 대하여 다음과 같은 의견을 제출합니다.

<대한민국 인공지능 행동계획>은 인공지능 윤리원칙의 제정 취지로 “윤리기준의 국제적 정합성을 재점검”하고 “개발·이용 사업자가 점검할 수 있는 체계적 기준”을 마련할 필요가 있다고 짚었다. 인공지능 윤리원칙이 다양한 주체의 의견을 정책 과정에 반영하여 사회적 합의를 기반으로 할 필요가 있다고도 밝혔다. 이에 AI 시민행동은 인공지능 윤리원칙(안)이 ▲국제 기준으로서 인권기반 접근의 반영, ▲인공지능기본법에 대한 윤리적 금지선 보완, ▲영향 받는 사람의 알 권리와 참여권 보장, ▲차별금지 원칙의 명시와 차별의 시정, ▲인공지능사업자의 책임성을 명시할 것을 요구한다. 특히「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(인공지능기본법)」이 올해 이미 시행 중인 상황에서, 인공지능 윤리원칙이 국제적 정합성을 제고하고 사업자의 바람직한 준수 기준을 제시하기 위해서는 무엇보다 영향 받는 사람의 피해 위험을 방지하고 구제 가능성을 높여야 할 것이다.

첫째, 국제 기준으로서 인공지능에 대한 인권기반 접근을 반영해야 한다. 우리 정부가 2020년 인공지능 윤리기준을 수립할 때는 “구속력 있는 ‘법’이나 ‘지침’이 아닌 도덕적 규범이자 자율규범으로” 기업 자율성을 존중하겠다는 지향을 밝혔던 바 있다. 이와 같은 기업 편향으로 인하여 당시에 국제적으로 널리 인정되고 있는 인공지능 기준인 “견고성 원칙”을 기업의 요구로 삭제하였고, 시민단체가 요구한 '투명성 원칙 강화'는 "학계·기업 다수의견과 충돌되어

미수용"하는 등 기업의 입장에 치우친 윤리원칙으로 국제적 수준에도 미달할 수 밖에 없었다(2020.12. 23. 제19차 4차산업혁명위원회 심의안건 제2호).

한편, 유엔 인권기구가 반복해서 강조하는 국제 기준은 "인공지능에 대한 인권기반 접근(Human Rights-Based Approach to AI)"이다. 2018년 유엔 의사표현의자유 특별보고관은 "머신러닝의 적응성으로 인해, 사람이 인공지능시스템의 목표와 결과를 정의하는 과정에서 점차 배제될 수 있다 ... 이는 인공지능의 투명성, 책임성 및 효과적인 구제수단에 대한 접근을 어렵게 만들 수 있다"고 우려하면서 이를 해결하기 위해서 국가와 기업이 인권영향평가 등 인권 침해 예방·완화·시정하고 국제인권법을 준수하는 인권기반 접근을 취해야 한다고 제안하였다. 다만 당시 민간과 공공 부문이 인권법 준수보다 '윤리' 규범을 내세우는 배경에 "인권기반 규제에 대한 저항"이 있다고 우려하였다(2018. 8. 29. A/73/348). 2020년 유엔 사무총장은 인공지능에 대한 인권기반 접근을 "사람을 권리주체 개인으로 대우하고, 권리를 행사할 수 있는 역량을 강화하며, 법적·제도적 환경을 조성하여 인권침해를 구제하는 접근법"이라고 소개하였다(2020. 3. 4. A/HRC/43/2). 2025년에는 유엔인권최고대표 사무소 역시 <기업과 인권 이행원칙>을 비롯한 국제인권규범을 인공지능 기업 또한 준수하여야 함을 강조하였다(2025. 6. 16. A/HRC/59/32.).

또한 지난 2024년 세계 최초의 인공지능 국제협약으로 공개되어 2026년 7월 현재 세계 21개국이 참가하고 있는 <인공지능과 인권, 민주주의 및 법치주의에 관한 유럽평의회 기본 협약>은 인권, 민주주의, 법치주의 준수를 명시적으로 요구하고 있다. 그러나 인공지능 윤리원칙은 3대 가치와 6대 원칙에서 '인권'을 명시적으로 언급하고 있지 않다. 다만 3대 가치에서 "인간의 존엄성"을 표방하고 있으나 여기에는 "인간의 생명·신체·건강·재산이 인공지능으로 인하여 침해되거나 훼손되어서는 아니 된다."고 서술하고 있을 뿐이다. 이는 2020년 인공지능 윤리기준이 인권보장을 제1의 핵심요건으로 명시한 것과 대비된다. 유럽연합(EU) <신뢰할 수 있는 AI 가이드라인>도 존엄·자유·평등·연대 등 기본권에 토대를 두고 인간 자율성 존중, 위해 방지, 공정성, 설명가능성의 네 윤리원칙을 도출하고 있다. 이처럼 인공지능 윤리원칙 또한 규범적 기준점으로서 기본권·인권을 명시적으로 천명할 필요가 있다. 따라서, "인간의 존엄성" 가치 부분은 세계인권선언, 여성차별철폐협약(CEDAW) 등 국제인권법이 인정하는 인권의 보호와 존중, 그리고 인권 침해를 구제해야 할 기업의 책임을 언급하여야 한다.

둘째, 인공지능기본법에 대한 윤리적 금지선을 보완하여야 한다. 인공지능기본법이 시행 중인 상황에서 윤리원칙은 법률에 포함되지 못한 인공지능 위험성을 규율하는 내용을 보완하는 것이 마땅하다. EU 신뢰할 수 있는 AI 가이드라인에서는 법은 항상 기술 발전의 속도를

따라가지 못하기에 윤리가 단순히 법을 준수하는 것을 넘어 법률이 기술 발전 속도를 따라가지 못해 발생하는 ‘입법 공백’을 선제적으로 보완해야 한다고 지적한 바 있다. 특히 현행 인공지능기본법은 금지하는 인공지능이 없기 때문에, 예를 들어 국제적으로 금지되고 있는 완전자율살상무기의 개발과 이용이나, 인공지능 결과물이 위헌 위법적으로 직접적이고 간접적인 차별을 낳는 경우에 대해서도 규율하고 있지 않다. 또한 장애와 연령 등 취약성을 공격하는 인공지능, 정치적 의견과 노동조합 소속 여부를 추론하는 인공지능, 공공장소에서 실시간으로 사람의 얼굴이나 동작을 식별하고 추적하는 인공지능, 직장과 학교에서의 감정인식 등 인권적으로 용인할 수 없는 수준의 인공지능 또한 금지하고 있지 않다. 따라서 인공지능 윤리원칙은 합법성의 공백을 보완하는 ‘윤리적 금지선(Ethical Red Line)’을 명문화하여야 한다. 따라서, 인공지능 윤리원칙의 “사회의 공공선” 가치 부분은 국제 인권 기준이 금지하고 나아가 인권과 양립할 수 없는 위험을 야기하는 인공지능을 예시적으로 열거한 후 그 개발과 이용에 대하여 금지할 것을 권장해야 한다.

셋째, 영향 받는 사람의 알 권리와 참여권을 보장하고 차별을 적극적으로 시정하여야 한다. 사회 각 분야에서 인공지능의 활용이 확산될수록 고용, 교육, 보건의료, 사회복지 등 삶과 노동에서 중대한 결정 또한 인공지능에 기반하여 이루어질 수 있다. 이때 인공지능은 그 결정의 영향을 받는 시민과 노동자의 안전과 인권에 중대한 영향을 미치거나 심지어 인권 침해적이거나 차별적인 결과를 초래할 수도 있다. 따라서 인공지능을 개발하고 활용하는 기업은 그 영향 받는 사람의 인권 침해를 구제하는 등 책임성을 갖출 필요가 있다. 인공지능 윤리원칙(안)은 부분적으로 인간의 감독, 설명 및 이의제기 등을 명시하고 있으며, 이는 인공지능기본법에서 이미 포함하고 있는 내용을 크게 벗어나지 않는다. 다만 현행 인공지능기본법은 인공지능 사업자가 영향 받는 사람을 보호해야 하는 책무와 절차를 명시적으로 규정하고 있지 않다. 이로 인하여 **2026. 1. 5.** 감사원이 지적한 바처럼 영향 받는 사람에게 발생하는 피해에 대한 보호 면에서 법률적 공백이 존재하는 상황에서 이러한 원칙 제시에는 의미가 없는 것이 아니다. 그러나 윤리원칙(안)의 ‘투명성’ 원칙은 ‘프라이버시, 영업비밀 보호 등 다른 정당한 가치와의 균형 속에서 구현되어야 한다’고 단서를 달았다. 하지만 윤리원칙에서 ‘영업비밀 보호’를 투명성을 제한하는 사유로 명시하는 것은 매우 부적절하며, 이러한 접근은 투명성의 실현을 기업의 영업비밀 보호를 넘어서지 못하고 형해화하는 결과를 낳을 우려가 있다.

투명성은 인공지능 사업자가 영향 받는 사람을 보호하고 구제하기 위하여 취해야 하는 책임성의 일환이라는 점을 상기할 필요가 있다. 의사표현의 자유 특별보고관이 앞서 지적하였듯이, 인공지능 사업자는 충분한 투명성, 책임성을 갖추고 그 영향을 받는 사람에 대하여 효과적인 구제수단을 제공하여야 한다. 특히 유럽평의회 기본협약은 규제기관에 대한

정보 제공을 요구하고 있으며, 유럽연합 인공지능법은 직장에 도입되는 인공지능에 대하여 사업자가 노동자와 노동조합에게 정보를 제공할 것을 요구한다. 따라서 인공지능 윤리원칙의 “투명성” 원칙은 영향 받는 사람에 대하여 사전적으로 충분한 정보를 제공하고, 그 피해 구제를 집행하여야 하는 규제기관에 대한 충분한 정보를 제공해야 할 책무 또한 명시하여야 한다. 나아가 “사회의 공공선” 가치 부분은 인공지능의 개발, 도입, 활용의 전 과정에서 영향을 받는 시민·노동자가 해당 정책 결정에 의견을 제시하고 참여할 수 있는 ‘민주적 인공지능 거버넌스’를 보장해야 한다.

넷째, 인공지능 윤리원칙(안)의 “공정성·포용성” 원칙은 데이터와 알고리즘의 편향에 대해 “지속적인 점검과 개선이 이루어져야 한다”는 추상적이고 소극적 권고에 그치는 것은 문제이다. 다양성, 공정성, 포용성의 의미를 너무 협소한 관점으로 서술하고 있다. 각각의 정의를 차별금지와 평등의 관점에서 서술해야 한다. 핵심문제는 성별, 장애, 인종, 연령 등에 대한 역사적·구조적 차별이 인공지능 시스템을 통해 재생산·고착화될 수 있다는 점이다. ‘포용’이라는 복지적 접근만으로는 이러한 문제를 해결할 수 없으며, 헌법상 평등권과 국제인권규범에 따라 ‘비차별(차별금지)’을 원칙 수준에서 명확히 분리하여 명시하여야 한다. 아울러 윤리원칙(안)이 인공지능 기술의 접근권을 포용성과 동일시하여 “기술적 소외가 사회적 배제로 이어지지 않도록 한다”는 디지털 격차 해소의 차원으로만 접근한 것 역시 문제이다. 직접적이고 간접적인 차별적 결과가 사람에게 영향을 미친 경우, 인공지능사업자는 그 시정과 구제에 나서는 것이 마땅한 윤리적 책임이고 이와 같은 윤리적 책임을 명시하여야 한다.

다섯째, 인공지능 윤리원칙은 인공지능사업자의 책임성을 명시하여야 한다. 2020년 윤리기준은 10대 핵심요건으로 “책임성”을 명시하였고, “발생할 수 있는 피해를 최소화”하고 설계·개발·서비스 제공자·사용자 간의 책임소재를 명확히 해야 한다고 서술하였다. 국제적으로는 EU 신뢰할 수 있는 AI 가이드라인에서 7대 요건으로 책임성을 제시하고 있으며, 책임성은 로그·문서화·감사가능성·거버넌스 체계와 접근가능한 구제까지 포함하는 항목이다. OECD 또한 2024년 인공지능 원칙 개정에서 추적가능성·위험관리 조항을 책무성 원칙으로 이관하여 그 위상을 강화한 바 있다. 그러나, 인공지능 윤리원칙의 6대 원칙은 EU 7대 요건과 사실상 일대일로 대응하면서 유독 책임성만 삭제하였으며, 2020년 윤리기준 및 OECD 원칙과 비교하여 보았을 때에도 책임성을 삭제한 것을 합리적으로 납득할 수 없다.

책임의 귀속을 다루는 원칙이 없으면 피해 발생시 책임 소재와 구제의 근거를 윤리원칙에 포함하지 않겠다는 의미나 다를 바 없다. 그러나 인공지능 행동계획은 윤리원칙 고도화의 검토 배경으로 “AI 기술이 일상과 산업 전반에 깊숙이 통합되면서 알고리즘의 책임성, 투명성, 공정성 등 사회적 신뢰를 확보하기 위한 윤리 체계의 고도화가 필수 과제로 부상하고 있다.”며

책임성을 중요한 배경 중 하나로 언급한 바 있다. 나아가 인공지능사업자의 충분한 책임성을 확보하기 위해서는 독립된 점검·감독을 함께 둘 필요가 있다. 따라서 인공지능 윤리원칙은 인공지능사업자가 사전적으로 피해를 최소화하고 사후적으로 구제의 책임을 지고 있음을 명시하는 책임성 원칙을 부활시키고 독립성을 보완하여야 한다. 끝.