# 인공지능 윤리원칙의 필요성 및 현황

## 1. 인공지능 윤리원칙

**[학습목표]**

- **AI 윤리가 부각된 배경에 대해 알고 설명할 수 있다.**
- **대표적인 국제적 인공지능(AI) 윤리 가이드라인에 대해 설명할 수 있다.**

**[학습내용]**

- **AI 윤리가 부각된 배경**
- **유럽의 신뢰 가능한 인공지능(AI)**
- **유네스코 인공지능(AI) 윤리권고**

### (1) AI 윤리가 부각된 배경

우리들의 실생활 속에 인공지능이 알게 모르게 도입되어 사용되고 있으며 앞으로 그 활용 가능성과 다양성이 더욱 확장될 것이다. 그렇지만 이로 인해 기존에는 생각하지 못하거나 생각할 필요가 없었던 문제나 주제들에 대해 윤리적 숙고의 필요성이 점점 크게 강조되고 있다. 따라서 '인공지능(AI) 윤리'는 인공지능(AI) 못지않게 우리에게 새롭게 다가오는 용어이지만, 낯설기보다는 오히려 친숙하게 느껴진다. 인공지능(AI)은 특수한 속성을 지니는 인공물로서 현상 차원에서 사회적 영향력을 지닌 행위 주체(Agent)에 상응하는 기능을 수행하기도 하지만, 그 작동 혹은 행위의 결과에 대하여 도덕적-법적 책임까지 질 수 있는 독립된 자율적 주체는 아니다. 인공지능(AI)은 그것의 '인공성'(artificiality), 즉 인간의 설계와 제작에 의하여 생성되고 속성이 결정된 산물임에도 불구하고 특이성, 특히 현상적으로 책임을 함축하는 행위 주체성과 자율성(agency or autonomy)을 지닌 것처럼 인식될 수 있는 이중성을 갖고 있으며, 이러한 이중성 때문에 인공지능(AI) 윤리의 필요성이 강조된다.

현대사회에서 다양하고 새로운 윤리에 대한 요청들이 제기되고 있다. 인공지능(AI) 윤리 분야에서도 로봇이나 자율주행차를 비롯하여 데이터와 같이 윤리학의 새로운 연구 주제와 대상이 나타나고 있다. 새로운 연구 대상에 대한 주제별 윤리학적 접근의 성과는 과학 기술 철학의 영역이지만 넓게 보면 실천윤리학의 영역일 것이다. 인공지능(AI) 윤리가 실천윤리학의 다른 영역과 차별화되는 특징은 인공지능(AI) 윤리가 '앞 북 치는' 윤리의 성격을 갖는다는 것이다. 철학의 미네르바처럼 윤리학도 대체로 어떤 문제가 발생하고 나서 그것에 대한 해결책을 찾으려는 시도에서 시작되기 마련인데, 인공지능(AI) 윤리는 예견적으로 발생 가능한 문제들에 대한 관심에서 출발하고 있으며 심지어 아직 기술적으로 가능하지 못한 것까지도 고려하고 논의하는 경우들이 있다.

또 다른 특징은 인간 중심적 윤리의 관점에서 벗어나 인간이 아닌 도덕적 행위자의 등장을 다루고 있다. 인공적 도덕 행위자(AMA: Artificial Moral Agent)는 자율적인 인간처럼 '도덕적 행위'를 판단해서 실행할 수 있는, 즉 도덕적 행위자라고 보기는 어려울 것이다. 그렇지만 주변 상황을 지각하고, 이를 근거로 해서 판단을 내리고 어떤 행위를 실행할 수 있는데 그 행위가 사람이나 사회에 도덕적인 측면에 영향을 줄 수 있는 행위라고 한다면 우리는 그러한 행위자를 도덕적 영역과 관련된 행위를 할 수 있는 '행위자' 즉, 도덕적 행위자로 간주할 수 있다. 인공지능(AI)이 인간과 같이 자유의지를 지닌 자율적 존재로 자리매김하지는 않겠지만, 적어도 현상적 차원에서 자율적 주체인 것처럼 행동할 수 있을 것이므로, 이런 맥락에서 '위임된 자율성' 혹은 '준 자율성(quasi-autonomy)'이라는 개념이 도출되기도 한다. 이 자율성은 인공지능(AI)에게 윤리적 사고와 판단 시스템을 부여하려는 시도가 이뤄지면서 더욱 강조되고 있다.

인간과 유사한 외형을 가진 로봇의 등장과 더불어 'the uncanny valley(불쾌한 골짜기, N. Tschoepe & M. Oehl, 2017 참조)'[1]에서 설명되는 것처럼 인간과 비슷해 보이는 로봇으로 인해 불안감, 혐오감, 두려움이 나타날 수도 있을 것이다. 그러나 이와는 정반대로, 마치 자율주행기술이 탑재된 자율주행차를 운행하기 전이나 운행을 처음 시도할 때 느끼는 불안감이, 반복된 자율주행기술의 경험으로 인해 'the overtrust peak (과신의 정점)'로 나타날 수 있는 것처럼 기술에 대한 지나친 믿음을 가져올 수도 있다. 지나친 믿음으로 인한 반성의 경험을 하게 되면 해당 기술에 대한 제한이나 조건들을 숙고하게 되고 과학 기술에 대한 긍정적, 부정적 반응을 통해 우리는 새롭게 다가오는 과학 기술의 변화와 이 변화가 주는 사회적 충격에 대한 숙고와 논의의 필요성을 깨닫게 된다.

본 차시에서는 지금까지 발표된 인공지능(AI) 윤리 가이드라인 중 가장 체계적인 것으로 2019년 EU가 발표한 '믿을만한(trustworthy) 인공지능(AI)과', 유네스코가 발표한 가장 최근에 그리고 가장 폭넓게 영향을 미칠 수 있는 '인공지능(AI) 윤리권고'를 중심으로 세계의 인공지능(AI) 윤리 동향을 살펴보고자 한다.

1) 인간이 인간이 아닌 존재를 볼 때, 그것이 인간과 더 많이 닮을수록 호감도가 높아지지만 일정 수준에 다다르면 오히려 불쾌감을 느낀다는 이론. Tschoepe, N., Reiser, J.E. & Oehl, M., Exploring the Uncanny Valley Effect in Social Robotics, HRI 2017 참조.

**(2) 유럽의 신뢰 가능한 인공지능(AI)**[2)]

2018년에 출범한 'EU의 인공지능(AI)에 대한 전문가그룹(High-Level Expert Group on Artificial Intelligence)'은 2019년 4월에 신뢰 가능한 인공지능(AI)의 틀을 법적 인공지능(AI), 윤리적 인공지능(AI), 건강한 인공지능(AI)로 규정하고, 4개의 윤리적 원칙, 7가지 요건 및 평가에 대해 규정하는 문서를 발표하였다.

이 문건은 법적인 인공지능(AI)에 대해서는 언급하고 있지 않으나, 4가지 윤리원칙과 7가지 요건들의 필요성을 제시하고 그 정당성을 설명하고 있다. 이러한 인공지능(AI) 윤리에 대한 가이드라인이 선언적 의미를 넘어서 윤리인증이나 표준화의 문제와 더불어 이미 법적, 경제적 중요성을 갖기 시작하였다는 것은 매우 중요한 변화이다.

유럽에서 강조되는 기본가치는 '인간의 존엄성 존중', '개인의 자유', '민주주의, 정의 그리고 법치에 대한 존중', '평등, 비차별 그리고 연대성', '시민의 권리'이다.

첫째, 인공지능(AI) 시스템은 인간의 육체적·정신적 통일성, 정체성의 개인적·문화적 의미, 인간의 본질적 욕구에 대한 만족을 존중하고 이에 기여하며 보호하는 방식으로 개발되어야 함을 천명하고 있다.

둘째, 인간은 자신의 삶을 자유롭게 결정해야 하는데, 이것은 주권 침해로부터의 자유를 포함한다. 다만 인공지능(AI)의 편익과 기회에 평등하게 접근할 수 있도록 정부나 비정부조직의 개입이 필요하다. 인공지능(AI)의 맥락에서 개인의 자유는 직·간접적으로 정당하지 못한 강요, 정신적 자율성과 정신 건강에 대한 위협, 정당하지 못한 감시, 기만, 공정하지 못한 조작의 완화를 요구한다.

셋째, 인공지능(AI) 시스템은 민주적 과정을 유지하고 촉진하는 데 기여해야 하고, 가치와 개인 선택의 다수성을 존중해야 한다. 인공지능(AI) 시스템은 민주적인 과정, 인간의 숙고나 민주적인 투표 체계를 약화시켜서는 안 된다. 인공지능(AI) 시스템은 법치주의의 기본적인 전제를 저해하는 방식으로 작동해서는 안되며 적절한 과정과 법 앞의 평등을 확고하게 만들겠다는 약속을 해야 한다.

넷째, 인공지능(AI)의 맥락에서 요구되는 평등은 인공지능(AI) 시스템이 불공정하고 편향된 산출을 하게 할 수 없다는 것을 포함한다. 예를 들어 인공지능(AI)를 교육시키기 위해 사용되는 자료는 가능한 한 포괄적이어야 하며 다양한 인구집단을 포함해야 한다. 또한 근로자, 여성, 장애인, 소수 민족, 아동, 소비자 또는 배제 위험에 처한 사람들

2) High Level Expert Group on Artificial Intelligence(2019), Ethics Guidelines for Trustworthy AI, pp. 8~11의 내용을 요약한 것임: https://ec.europa.eu/digital single-market/en/high-level-expert-group-artificial-intelligence 참조.

인공지능의 교육적 역할 및 윤리적 활용

과 같이 잠재적으로 취약한 개인과 집단에 대한 적절한 존중이 필요하다.

끝으로, 인공지능(AI) 시스템은 공공재 및 서비스 제공에서 정부의 규모와 효율성을 향상시킬 수 있는 실질적인 잠재력을 제공한다. 그렇지만 이와 동시에 시민의 권리가 인공지능(AI) 시스템에 의해 부정적인 영향을 받을 수 있으므로 보호되어야 한다.

위에서 제시한 기본적인 권리들을 토대로 하여 4가지 윤리원칙이 제안된다. 이러한 4원칙은 원칙 간의 위계 구조를 의미하지는 않지만 대체로 기본권에 대한 선언이나 자료에 나오는 순서를 반영하고 있음을 밝히고 있다.

우선 인간의 자율성 존중 원칙은 인간의 자유와 자율성에 대한 존중을 보장한다는 것이다. 인간과 인공지능(AI)의 상호작용에서 인간 자신이 자기 결정을 내릴 수 있어야 하고, 민주적 과정에 참여할 수 있어야 한다. 인공지능(AI) 시스템은 부당하게 인간을 억압, 강요, 기만, 조작, 조건 지우거나 몰아가서는 안 된다. 오히려 인공지능(AI) 시스템은 인간의 지적, 사회적 그리고 문화적 능력을 향상시키고 보완하며 강화하도록 설계되어야 한다. 인간과 인공지능(AI) 시스템의 기능 할당은 인간 중심적 설계 원칙을 지켜야 하며 인간의 선택에 적절한 기회의 여지를 남겨야 한다. 이것은 인공지능(AI) 시스템의 작업 과정에 대한 인간의 감시를 확실히 해야 함을 의미한다. 인공지능(AI) 시스템은 노동의 영역을 근본적으로 변화시키는데, 작업환경에서 인간을 도와주고 의미있는 작업을 만들어내는 것을 그 목적으로 삼도록 해야 한다.

해악 금지 원칙은 인공지능(AI) 시스템이 피해를 발생시키거나 악화시켜서는 안 되며, 인간에게 불리한 영향을 미쳐서는 안 된다는 것을 의미한다. 이 원칙은 인간의 존엄성뿐만 아니라 정신적, 육체적 통일성을 보호하는 것을 포함한다. 인공지능(AI) 시스템과 그 운영 환경은 안전(safe)하고 보호(secure)되어야 한다. 또한 인공지능(AI) 시스템은 기술적으로 견고해야 하며 악의적인 사용에 노출되지 않아야 한다. 사회적 약자는 보호를 받아야 하며, 인공지능(AI) 시스템의 개발, 배치 그리고 사용에서도 포함되어야 한다. 인공지능(AI) 시스템이 고용주와 피고용인, 사업체와 소비자, 정부와 시민 사이에 권력이나 정보의 비대칭성으로 인해 부정적인 영향을 유발하거나 악화시키는 상황에 특히 조심해야 한다. 금지되는 해악에는 자연환경이나 모든 생명체에 대한 해악도 포함된다.

인공지능(AI) 시스템의 개발, 배포 및 사용은 공정해야 한다는 공정성 원칙은 실질적 차원과 절차적 차원을 가지고 있다. 실질적 공정성 측면은 복리 후생 및 비용이 평등하고 정당하게 분배되고, 개인과 그룹이 불공정한 편견, 차별 및 낙인으로부터 자유로워지도록 보장한다. 불공정한 편견을 피할 수 있다면 인공지능(AI) 시스템은 사회적 공정성을 높일 수도 있다. 실질적 차원의 공정성 확보를 위해 교육, 재화, 서비스 그리고 기술에 대한 접근에서의 평등한 기회가 조성되어야 하며 인공지능(AI) 시스템의 사용으

로 사람들이 속거나 선택의 자유가 손상되어서는 안 된다. 이와 더불어 공정성은 인공지능(AI) 담당자들이 수단과 목적 사이의 비례 원칙을 존중하고, 경쟁적인 이해관계와 목적들 간의 균형을 맞추는 방법을 신중하게 고려해야 함을 의미한다. 공정성의 절차적 차원은 인공지능(AI) 시스템과 시스템을 운영하는 사람이 내린 결정에 대하여 이의를 제기하거나 효율적인 배상을 청구할 수 있는 능력을 포함한다. 이렇게 하기 위해서는, 결정에 대한 책임이 있는 실체를 찾아야 하고, 의사결정 과정이 설명될 수 있어야 한다.

설명 가능성의 원칙은 인공지능(AI) 시스템에 대한 사용자들의 신뢰를 구축하고 유지하기 위해 매우 중요한 원칙이다. 인공지능(AI) 시스템의 과정이 투명해야 하고, 인공지능(AI) 시스템의 능력과 목적이 공개적으로 밝혀져야 하며 결정들이 직접적, 간접적으로 영향을 받는 사람들에게 최대한 설명될 수 있어야 한다는 것을 의미한다. 이러한 정보가 없다면 내려진 결정에 대한 적절한 이의를 제기할 수 없다. 하나의 모델이 특정한 산출이나 결정을 생성한 이유(그리고 입력요인의 어떤 결합이 결정에 기여한 이유)에 대한 설명이 항상 가능한 것은 아니다. 이러한 경우를 '블랙박스' 알고리즘이라고 하며 특별한 주의를 요한다. 이런 상황에서는 시스템이 기본적인 권리를 존중한다면, 다른 설명 조치(예를 들면 추적 가능성, 감사 가능성, 시스템의 능력에 대한 투명한 의사소통)가 필요할 수도 있다. 설명 가능성이 필요한 정도는 산출이 잘못되거나 정확하지 못할 경우 결과의 심각성과 맥락에 의존한다.

이러한 원칙들이 인간의 기본권에서 도출되었다 하더라도 원칙들 간의 충돌은 불가피하다. 그래서 강조되는 것이 바로 민주적 참여, 적절한 과정과 절차의 구성, 공개적인 정치 참여, 긴장을 다룰 수 있는 책임 있는 숙고 방법들이다. 이를테면 해악 금지 원칙과 인간의 자율성 존중 원칙이 상충 될 수 있다. 범죄 예방을 위해 인공지능(AI) 시스템을 활용하는 경우를 생각해보자. 인공지능(AI) 시스템을 활용한 감시 기술이 범죄를 줄일 것으로 기대되지만 동시에 개인의 자유와 프라이버시를 침해할 수 있다. 공정성 원칙과 설명 가능성 원칙의 경우도 충돌 가능성이 있다. 설명 가능성을 기업에 요구하는 것이 기업의 영업비밀을 드러내는 부담감으로 인하여 불공정한 요구로 들릴 수도 있기 때문이다.

### (3) 유네스코 인공지능(AI) 윤리권고[3)]

2021년 11월 23일 프랑스 파리 유네스코 본부에서 개최된 제41차 유네스코 총회에서 '유네스코 인공지능 윤리권고'가 채택되었다. 이 권고의 주요 내용을 살펴보면 다음과 같다.

우선 이 권고에서 강조되는 가치는, 첫 번째 인간의 권리와 기본적 자유 그리고 인간의 존엄성이다. 두 번째 환경과 생태계의 번영, 세 번째 다양성과 포괄성의 보장이고, 네 번째 평화롭고 정의로운 서로 연결된 사회에서 살기이다. 이 윤리권고는 인공지능(AI) 시스템의 전체 생애 주기 안에서 가치가 인정되고, 보호되며 촉진되어야 함을 주장하고 있다. 초안의 상충하거나 수준이 다른 가치들의 문제를 포괄적인 범주로 재배열하여 보다 설득력 있게 제시하고는 있지만, 가치들의 수준 층 차의 문제점이 해소된 것처럼 보이지는 않는다.

이 윤리권고는 비례성과 해악 금지, 안전과 보안, 공정성과 비차별, 지속가능성, 사생활의 권리와 데이터 보호, 인간의 감시와 결정, 투명성과 설명 가능성, 책임성, 공적 의식과 리터러시, 다양한 관계자와 적응적 거버넌스와 협력의 원칙들을 제안하고 있다. 비례성 원칙은 인공지능(AI) 시스템이 정해진 정당한 목적의 실현에 있어서 적절해야 함을 의미하며, 지속가능성 원칙은 인공지능(AI) 시스템이 지속가능한 발전의 목표를 포함하여 지속가능성을 보장해야 함을 의미한다. 인간에 의한 감시의 원칙은 개인의 감시뿐만 아니라 대중의 감시 가능성도 포함하고 있다. 최근 알고리즘 기계학습 과정이 설명 가능하지 않고 의미의 변화도 겪고 있는데, 설명 가능성 원칙은 인공지능(AI) 시스템의 산출이 이해 가능하고, 추적 가능하면서 사회적 맥락에 적절해야 함을 투명성의 원칙과 결합하여 제시하고 있다. 공적 의식과 리터러시 원칙은 인공지능(AI) 기술과 데이터의 가치에 대한 대중들의 의식과 이해가 공적 교육, 시민교육, 인공지능(AI) 윤리 교육, 인공지능(AI) 리터러시 등을 통해 촉진되어야 함을 의미한다.

물론 이러한 가치와 원칙의 제시는 선언적 의미가 강하고, 세계 각국의 정치적, 경제적 수준의 다양성으로 인해 일관되거나 체계적이기 어렵다는 한계를 갖고 있다. 특히 가치들 간의 상충이나 원칙들 간의 충돌 가능성 해결 등에 대한 논의가 열려 있다고 해야 할 것이다. 그럼에도 불구하고 정책에 대해 매우 강력한 권고를 하고 있다는 점에서 그 실천적 의미가 있다고 볼 수 있다. 윤리 영향평가, 윤리적 거버넌스와 관리, 데이터 정책, 개발과 국제적 협력, 환경과 생태계, 성, 문화, 교육과 연구, 공동체와 정보, 경제와 노동, 건강과 사회적 웰빙 분야의 정책 이슈들은 지금 그리고 앞으로 인공지능 시대에 반드시 고려해야 할 정책적 이슈와, 사회가 앞으로 심각하게 고민해야 할 문제들을 제시하고 있다.

비록 다양한 이해관계 당사국들의 논의를 종합하다 보니 체계적이고 윤리학적인 담론의 수준에서 부족한 부분들이 보이지만, 그래도 어느 정도 국제적 구속력을 가진 유네스코 권고라는 점에서 규범적 의미와 구속력을 가진다고 볼 수 있다.

3) https://unesdoc.unesco.org/ark:/48223/pf0000381137 참조: 이에 대한 한국어 해설서는 이상욱(2021), 유네스코 인공지능 윤리 권고 해설서, 유네스코한국위원회가 있음.

인공지능의 교육적 역할 및 윤리적 활용

인공지능의 교육적 역할 및 윤리적 활용

## 2. 국가 인공지능 윤리기준과 교육분야 인공지능 윤리원칙의 의미

**[학습목표]**

- **국가 인공지능 윤리기준의 의미를 알고 적용할 수 있다.**
- **교육분야 인공지능 윤리원칙의 의미를 알고 적용할 수 있다.**

**[학습내용]**

- **국가 인공지능 윤리기준의 의미**
- **교육분야 인공지능 윤리원칙의 의미**

### (1) 국가 인공지능 윤리기준의 의미

우리나라는 최근에(2020.12.23.) 사람이 중심이 되는 인공지능(AI) 윤리기준을 발표하였다. 2019년까지 'Seoul PACT'라고 불리는 4대 원칙 38개 세부 지침으로 구성된 지능정보사회 윤리 가이드라인을 한국정보화진흥원에서 제시하였으며, 3가지 기본가치와 5대 실천원칙으로 구성된 인공지능·로봇(의 개발과 이용)에 대한 윤리 가이드라인을 로봇산업진흥원에서 시험적으로 제안했다.[4] 2020년 12월 23일에는 과학기술정보통신부를 중심으로 바람직한 인공지능 개발·활용 방향을 제시하기 위해 "사람이 중심이 되는 「인공지능(AI) 윤리기준」"을 발표하였다. 새로 발표한 인공지능의 윤리기준은 3대 기본원칙과 10대 핵심 요건으로 이루어져 있으며 부록에서 인공지능의 지위에 관해 다루고 있다.

3대 기본원칙은 인간 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙이다. 그 내용은 다음과 같다.


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 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 행복의 가치를 추구한다.
- 인공지능은 지능정보사회에서 소외되기 쉬운 사회적 약자와 취약계층의 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.

4) 변순용, "AI로봇의 도덕성 유형에 근거한 윤리인증 프로그램(ECP) 연구", 윤리연구, 2019, p.77.

- 인류의 삶과 번영을 위한 인공지능 개발 및 활용을 장려하여 진흥해야 한다.

인간 존엄성의 원칙은 인간을 위해 개발된 기계제품과 인간은 교환 불가의 절대성을 가지고 있으며 인공지능이 인간의 생명이나 정신적, 신체적 건강에 해가 되지 않고 안전하게 개발되어야 한다는 의미이다. 이는 과거에는 인간의 존엄성이 타인이나 제도에 의해서만 침해받아 왔지만, 앞으로는 인공지능(AI)에 의해 침해받을 수 있다는 것을 의미한다. 이 원칙은 인간의 존엄성이 타인뿐만 아니라 인공지능(AI)에 의해서도 위협받거나 억압되고 침해받지 않기 위한 원칙이다. 인공지능(AI)에 의해 인간이 분류되거나 점수화되거나 하나의 데이터 또는 데이터 조건으로 취급될 수 있는 대상이 아님을 의미한다. 인공지능(AI) 시스템은 이러한 원칙을 지켜 인간의 본질적 욕구를 존중하고 기여하며 보호하는 방식으로 개발되어야 한다. 또한 이 원칙은 인공지능(AI)을 활용하는 개인 또는 정책 개발자에게 인공지능(AI)을 활용해 다른 사람의 존엄성을 침해하지 말아야 한다는 것까지 포함한다.

사회의 공공선 원칙을 준수하는 인공지능(AI)은 사회 공동체에 가능한 많은 혜택(행복, 안녕 등)을 제공할 수 있는 가치를 추구해야 한다. 인공지능(AI)은 활용하기에 따라 인공지능(AI)을 소유하고 활용할 수 있는 특정 계층에 독점될 수 있다. 이에 인공지능(AI)은 미래의 지식정보사회에서 소외되기 쉬운 사회적 약자, 취약 계층에게도 충분한 접근성을 보장하고 그들이 쉽게 활용할 수 있도록 개발되어야 한다. 이는 사회적, 국가적, 세계적 관점에서 인류 보편적 복지의 향상을 의미한다. 그리고 사회에서 인공지능(AI) 시스템을 통해 약화될 수 있는 민주주의 및 법치주의 시스템이 손상되어서는 안 된다는 것을 포함한다.

기술의 합목적성 원칙을 준수하는 인공지능(AI) 기술은 인류의 삶을 위한 기술이라는 목적에 부합되도록 개발 과정과 활용 모두 윤리적이어야 한다. 인공지능(AI) 기술이 인류의 가치와 윤리에 부합되지 않고 인류 발달에 기여하지 못하게 오·남용될 경우 인간은 큰 피해를 입게 될 것이다. 예를 들어 사람에게 편리하고 알맞은 서비스를 제공하기 위한 빅데이터 기반의 알고리즘은 편향성과 편견을 주입하거나 특정한 목적을 달성하기 위해 인간에게 정보를 선별하여 전달하는 식으로 오용될 수 있다. 이와 같이 인공지능(AI) 기술은 본래 개발된 목적에 맞는 합목적성을 갖춰서 활용되어야 한다. 이러한 기술의 합목적성의 원칙은 기술을 활용하거나 정책을 만드는 사람들이 갖추어야 할 합리적 도덕 역량이라고 할 수 있다.

인공지능의 교육적 역할 및 윤리적 활용

3대 기본원칙을 실현할 수 있는 세부 요건을 10대 핵심 요건이라고 하며 아래 그림과 같이 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성으로 이루어져 있다.


① 인권보장

- 인공지능(AI)의 개발과 활용은 모든 인간에게 동등하게 부여된 권리를 존중하고, 다양한 민주적 가치와 국제 인권법 등에 명시된 권리를 보장하여야 한다.
- 인공지능(AI)의 개발과 활용은 인간의 권리와 자유를 침해해서는 안 된다.

② 프라이버시 보호

- 인공지능(AI)을 개발하고 활용하는 전 과정에서 개인의 프라이버시를 보호해야 한다.
- 인공지능(AI) 전 생애주기에 걸쳐 개인 정보의 오용을 최소화하도록 노력해야 한다.

③ 다양성 존중

- 인공지능(AI) 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(AI) 기술 및 서비스에 대한 접근성을 보장하고, 인공지능(AI)이 주는 혜택은 특정 집단이 아닌 모든 사람에게 골고루 분배되도록 노력해야 한다.

④ 침해금지

- 인간에게 직간접적인 해를 입히는 목적으로 인공지능(AI)을 활용해서는 안 된다.
- 인공지능(AI)이 야기할 수 있는 위험과 부정적 결과에 대응 방안을 마련하도록 노력해야 한다.

⑤ 공공성

- 인공지능(AI)은 개인적 행복 추구뿐만 아니라 사회적 공공성 증진과 인류의 공동 이익을 위해 활용해야 한다.
- 인공지능(AI)은 긍정적 사회변화를 이끄는 방향으로 활용되어야 한다.
- 인공지능(AI)의 순기능을 극대화하고 역기능을 최소화하기 위한 교육을 다방면으로 시행하여야 한다.

⑥ 연대성

- 다양한 집단 간의 관계 연대성을 유지하고, 미래세대를 충분히 배려하여 인공지능(AI)을 활용해야 한다.
- 인공지능(AI) 전 주기에 걸쳐 다양한 주체들의 공정한 참여 기회를 보장하여야 한다.
- 윤리적 인공지능(AI)의 개발 및 활용에 국제사회가 협력하도록 노력해야 한다.

⑦ 데이터 관리

- 개인정보 등의 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(AI) 개발 및 활용과정에서 책임주체를 설정함으로써 발생할 수 있는 피해를 최소화하도록 노력해야 한다.
- 인공지능(AI) 설계 및 개발자, 서비스 제공자, 사용자 간의 책임소재를 명확히 해야 한다.

⑨ 안전성

- 인공지능(AI) 개발 및 활용 전 과정에 걸쳐 잠재적 위험을 방지하고 안전을 보장할 수 있도록 노력해야 한다.
- 인공지능(AI) 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(AI) 활용 상황에 적합한 수준의 투명성과 설명 가능성을 높이려는 노력을 기울여야 한다.
- 인공지능(AI) 기반 제품이나 서비스를 제공할 때 인공지능(AI)의 활용 내용과 활용 과정에서 발생할 수 있는 위험 등의 유의사항을 사전에 고지해야 한다.

추가적으로 부록에서 인공지능(AI)의 지위는 인간을 대체하거나 교환할 수 없으며, 인간을 위한 수단으로 명시하고 있으나 인간종 중심주의나 인간 이기주의를 표방하는 것은 아니다. 인공지능(AI)을 수단으로 둘 수 있는 이유는, 지각력을 가지고 스스로를 인식하고 사고하여 행동할 수 있는 독립된 인격의 강인공지능(AI)이 아니라 도구로 사용되는 인공지능(AI)을 상정하기 때문이다. 따라서 우리나라에서는 인공지능(AI)의 발달이 인간의 권리와 생활에 큰 영향을 줄 것으로 생각하고 이를 보호하기 위한 기준을 만들고 있음을 알 수 있다.

위의 3대 기본원칙과 10대 핵심요건은 인공지능(AI)을 개발하고 활용하는 과정에서 인간에게 피해를 주지 않도록 하는 것이 핵심이다. 우리나라는 윤리기준을 발표하면서 법이나 지침이 아닌 도덕적 규범이자 자율 규범으로 하였으며 특정 분야에 제한되지 않는 범용성을 가진 일반원칙이다. 즉 모든 인간이 인공지능(AI)을 개발, 활용하기 위해 지켜야 하는 기본 요건이다. 이는 인공지능(AI)을 개발하고 활용함에 있어서 윤리적인 고려가 반드시 필요함을 의미한다.

### (2) 교육분야 인공지능(AI) 윤리원칙의 의미[5)]

그동안 국내외에서 발표된 인공지능(AI) 윤리기준(원칙)은 대부분 범용적이고 개발자(공급자) 중심으로 규범화되어, 교육 현장에 바로 적용하기에는 제한적일 수밖에 없었다. 교육에 활용되는 인공지능(AI)은 학습자의 성장에 미치는 영향이 작지 않을 것이고, 교육 현장 및 수업에도 변화를 가져올 것으로 예상되어 교육분야에서 인공지능(AI) 활용에 필요한 윤리적 기준에 대한 필요성이 제기되고 있다. 실제로 미국의 알트스쿨의 사례에서와 같이 인공지능(AI) 기반 스마트기기 등을 활용하여 학생 맞춤형 서비스를 제공(교사 역할 배제)하자 학생의 학업성취도 저하 및 문맹률 증가의 문제가 발생하였다. 이는 기술의 활용도 중요하지만 교사의 역할, 학생과의 소통이 교육의 중요한 기제로 작용하고 있음을 확인할 수 있는 계기가 되었다.

교육분야 인공지능(AI) 윤리원칙은 교육당사자 및 관계자가 교육현장, 개발현장, 정책현장 등에서 인공지능(AI) 활용교육, 교육용 인공지능(AI) 개발, 인공지능(AI) 관련 정책마련 등을 위해 교육기관 및 행정기관에서 활용하는 인공지능(AI)이 윤리적으로 개발, 활용될 수 있도록 자발적으로 실천, 준수하는 자율규범이라고 밝히고 있다. 교육분야 인공지능(AI) 윤리원칙의 전체 틀은 아래와 같다.

5) 교육부(2022), 사람의 성장을 지원하는 교육분야 인공지능 윤리원칙 참조.

인공지능의 교육적 역할 및 윤리적 활용


인공지능의 교육적 역할 및 윤리적 활용

국가수준 인공지능(AI) 윤리기준에서 최고의 가치로 제시되는 인간성과 관련하여 교육분야 인공지능(AI) 윤리원칙에서는 인공지능(AI)이 사람의 성장을 지원해야 함을 최고의 가치로 설정하고 있다. "교육분야 인공지능(AI)이 사람의 전 생애에 걸쳐 전인적 성장 지원을 최고 가치로 삼으며, 사람의 인격을 존중하고 개성을 중시하여 사람의 능력이 효과적으로 발휘될 수 있도록 제공되어야 함"[6]이 강조되고 있다.

이러한 교육분야 인공지능(AI) 윤리원칙을 통한 향후 실천 계획은 우선, 인공지능(AI) 윤리 커리큘럼을 개발하고, 교육당사자에게 인공지능(AI) 윤리교육 프로그램을 제공한다. 특히, 경제적·사회적 취약계층 학생 및 장애 학생 등의 인공지능(AI) 윤리교육에 대한 접근을 보장한다. 둘째, 교사의 인공지능(AI) 역량 강화인데, 교수자 연수를 통해 인공지능기술의 윤리적 활용에 필요한 교수자 역량 제고를 위해 교사의 인공지능(AI) 교육역량 체계 정립과 교원대상 원격연수 콘텐츠 개발을 적극 지원한다. 셋째, 교육분야 인공지능(AI) 활용과정에서 발생 가능한 윤리 이슈를 선제적으로 발굴 하고, 학술 연구를 지원한다. 끝으로, 교육분야 인공지능(AI)의 윤리적 활용을 위해 정부·기관·학교 등과의 협력체계를 구축하고, 현장의 문제해결을 위한 도구 개발 및 서비스 등을 지원한다.

6) 교육부(2022), 교육분야 인공지능 윤리원칙, p.7.