‘원칙에 입각한 인공지능’ 보고서 개요

지난 몇 년 동안, 기술 정책과 관련있어 보이는 모든 조직이 AI에 대한 일련의 원칙을 만들거나 승인했다. 윤리적이고, 권리존중적(rights-respecting)이며, 사회적으로 이로운 AI에 대한 지침이 기반 기술과 함께 빠르게 발전함에 따라, 개별적으로 또 맥락적으로 이들을 이해해야 할 절박한 필요성이 있다. 이를 위해 우리는 36건의 주요 AI 원칙의 내용을 분석하고 그 과정에서 분야별 표준의 출현을 시사하는 주제와 관련된 경향을 발견하였다.

각각의 원칙은 AI의 거버넌스에 대한 비전을 제시한다는 동일한 기본 목적을 추구하지만, 우리의 데이터셋에 있는 원칙은 다양하다. 원칙들은 목표 대상, 구성, 범위 및 깊이에서 다양하다. 원칙들은 중남미, 동아시아와 남아시아, 중동, 북미, 그리고 유럽에서 왔고, 문화적 차이는 의심할 여지 없이 내용에 영향을 미친다. 그러나 아마도 가장 두드러진 특징은 원칙들이 정부, 정부간 조직, 기업, 전문적 협회, 시민사회단체, 멀티스테이크홀더 이니셔티브 등 다양한 주체에 의해 작성되었다는 점이다. 시민사회와 멀티스테이크홀더가 만든 원칙은 목표 과제를 설정하거나 진행 중인 논의를 위한 장을 마련하는 역할을 할 수 있다. 각 국의 정부가 만든 원칙은 종종 전반적인 국가 AI 전략의 일부로 제시된다. 많은 민간 부문의 원칙은 저작 조직 내부의 AI 기술 개발과 사용을 통제하고 고객과 규제당국을 포함한 다른 이해관계자들에게 조직의 목표를 전달하기 위한 의도로 보인다. 수많은 축에 걸친 다양성의 범위를 감안할 때, AI 원칙에 대한 면밀한 연구가 공통 주제를 밝혀낸 것은 더욱 놀라운 일이다.

연구 결과의 첫 번째 중요한 측면은 8가지 핵심 주제 그 자체이다.

- 프라이버시. 이 주제에 따른 원칙은 기술 시스템 개발을 위한 데이터 사용에 있어서 그리고 영향을 받는 사람들에게 그들의 데이터 및 데이터에 기반한 의사 결정에 대한 권한((agency)을 제공함으로써 AI 시스템이 개인의 프라이버시를 존중해야 한다는 생각을 나타낸다. 프라이버시 원칙은 데이터셋에 있는 원칙들의 97%가 갖고 있다.
- 책임성. 이 주제는 AI 시스템의 영향에 대한 책임성이 적절히 분산되고 충분한 구제책이 제공되도록 보장하는 메커니즘의 중요성에 관한 원칙을 포함한다. 책임성 원칙은 데이터셋 원칙들의 97%가 갖고 있다.
- 안전과 보안. 이러한 원칙은 AI 시스템이 안전하고 의도한 대로 작동하며 부정침입을 막을 수 있게 보안이 철저해야 한다는 요구를 나타낸다. 데이터셋에 있는 원칙의 81%가 안전과 보안 원칙을 갖고 있다.
- 투명성과 설명가능성. 이 주제에 따른 원칙은 AI의 작동을 이해할 수 있는 결과물로 완전히 변환하는 것과 AI가 어디서, 언제, 어떻게 사용되는지에 대한 정보를 제공하는 것 등 감독이 가능하도록 AI 시스템을 설계하고 구현할 것을 요구한다. 투명성과 설명가능성 원칙은 데이터셋 원칙들의 94 %가 갖고 있다.
- 공정성과 비차별. 이미 전 세계적으로 AI 편향에 대한 우려가 개인에게 영향을 미치고 있는 가운데 공정성과 비차별 원칙은 AI 시스템이 공정성을 극대화하고 포괄성을 촉진하도록 설계되고 사용될 것을 요구하고 있다. 공정성과 비차별 원칙은 데이터셋 원칙들의 100%가 갖고 있다.
- 기술의 인적 통제. 이 주제에 따른 원칙은 중요한 결정들은 여전히 인간에 의한 검토의 대상이 될 것을 요구한다. 기술의 인적 통제 원칙은 데이터셋에 있는 원칙들의 69%가 갖고 있다.
- 전문적 책임. 이러한 원칙은 AI 시스템의 개발 및 배치에 관여하는 개인이 시스템의 영향에 미치는 중요한 역할을 인식하고, 적절한 이해관계자와 협의하고 장기적인 영향을 계획하는 데

있어 그들의 전문성과 청렴성을 촉구한다. 전문적 책임 원칙은 데이터셋에 있는 원칙들의 78%가 갖고 있다.

- 인간의 가치 옹호. 마지막으로, 인간 가치 원칙은 AI가 헌신하는 목적과 그것이 구현되는 수단은 우리의 핵심 가치와 일치해야 하며 인류의 행복을 전반적으로 증진시켜야 한다고 명시한다. 인간의 가치 옹호 원칙은 데이터셋의 원칙 중 69%가 갖고 있다.

연구 결과의 두 번째, 그리고 어쩌면 더 놀라운 측면은, 최근의 문헌들이 위 여덟 가지 주제를 모두 다루는 경향이 있다는 것인데, 이는 적어도 이 원칙들의 개발을 책임진 공동체에서 원칙에 입각한 AI에 대한 대화가 수렴되기 시작했음을 암시한다. 따라서, 이들 여덟 가지 주제는 AI 윤리 및 거버넌스에 대한 원칙에 기반한 접근법의 "규범적 핵심(normative core)"을 대표할 수 있다.[1]

그러나 우리는 독자들이 어떤 원칙이 핵심 주제들을 더 많이 포괄한다고 해서 더 낫다고 추론하지는 않길 바란다. 맥락(Context)이 중요하다. 각 원칙은 문화적, 언어적, 지리적, 조직적 맥락에서 이해되어야 하며, 어떤 주제는 다른 주제보다 특정 맥락과 청중에 더 관련이 있을 것이다. 더구나 원칙은 거버넌스의 종점이 아니라 출발점이다. 하나의 원칙 그 자체만으로는 강한 설득력을 갖기 어렵다. 원칙의 영향은 예를 들어 관련 정책(예: AI 국가 계획), 법률, 규제뿐 아니라 실무 관행과 일상 등 더 큰 거버넌스 생태계에 그것이 어떻게 자리잡고 있는지에 달려 있을 것이다.

AI 시스템의 영향에 대해 상당한 잠재적 관련성을 가진 기존 거버넌스 규범 중 하나는 국제인권법이다. 학자, 옹호자, 그리고 전문가들은 AI 거버넌스와 인권에 관한 법과 규범 사이의 연관성에 점점 더 주의를 기울여왔고[2], 우리는 우리가 연구한 원칙들 사이에서 이러한 관심이 미친 영향을 관찰했다. 원칙의 64%는 인권에 대한 언급을 포함했으며 다섯 건의 원칙은 그들의 전반적인 노력을 위한 틀로 국제인권을 채택했다. 인권의 해석과 보호를 위한 기존의 메커니즘은 개개인의 사례와 결정에 원칙을 적용할 때 유용한 조언을 제공할 수 있는데, 원칙의 적용은 "프라이버시"나 "공정성"과 같은 기준의 정밀한 판정을 필요로 할 뿐만 아니라, 하나의 원칙 안에서 개별 원칙이 팽팽히 맞서는 복잡한 상황에 대한 해결책을 필요로 하기 때문이다.

“원칙에 입각한 인공지능” 백서에 수록된 36건의 원칙은 특히 두드러지거나 영향력이 큰 원칙에 초점을 맞추어 다양하게 선별되었다. 위에서 언급한 바와 같이, 다양한 분야, 지역, 및 접근법이 나타났다. 우리의 주관적인 샘플링 방식과 윤리적, 권리존중적(ethical and rights-respecting) AI 분야가 여전히 매우 초창기라는 점을 감안할 때, 우리는 관점이 여기에 반영된 것 이상으로 계속해서 진화할 것으로 기대한다. 우리는 본 백서와 함께 제공되는 데이터 시각화가 윤리적, 권리존중적(ethical and rights-respecting) AI에 대한 대화를 진전시킬 수 있는 자원이 되기를 바란다.

---

[1] 우리가 발견한 두 가지 측면은 본 백서와 함께 제공되는 데이터 시각화(p. 8-9)에서 관찰할 수 있다.

[2] Hannah Hilligoss, Filippo A. Raso and Vivek Krishnamurthy, ‘It’s not enough for AI to be “ethical”; it must also be “rights respecting”’, Berkman Klein Center Collection (October 2018) https://medium.com/berkman-klein-center/its-not-enough-for-ai-to-be-ethical-it-must-also-be-rights-respecting-b87f7e215b97.