KAIST

KPC4IR
Korea Policy Center for the
Fourth Industrial Revolution

ISSUE PAPER No. 08

인공지능의 윤리/정책/사회 이슈

Ethical, Social and Policy Issues of AI

December, 2019

Contents

This report is written by:

Yang Jongmo

요 약

인공지능의 도입으로 인류 사회는 커다란 변혁의 길로 접어들고 있다. 인공지능의 사회 · 경제적 파급효과로 인한 여러 가지 변화, 특히 부작용은 인공지능에 대한 다양한 대응책을 강구하지 않을 수 없게 만들었다.

인공지능의 활용 과정에서 인공지능의 여러 가지 부정적인 측면이 노출되기 시작했다. 특히 인공지능의 불투명성, 편향성 때문에 인공지능 알고리즘의 적용 결과가 차별을 보이면서 사회적 문제로 대두되기도 하고, 자율주행자동차가 인공지능 알고리즘의 결함으로 인명사고를 일으키는 등의 문제가 발생하면서 안전성에 대한 우려도 커졌다. 이와 같이 인공지능이 초래하는 여러 가지 문제에 대한 해결방안으로는 문제 발생 후의 사후적인 피해 구제책도 있지만, 사전에 미리 예방할 수 있는 방안이 모색되어야 한다. 법적 규범에 의한 사전예방책도 있지만, 인공지능의 개발 단계에서부터 사용단계까지의 전 과정에서 개발자 등이 준수하여야 할 여러 가지 당위적 조치의무를 일률적으로 규정하는 방안이 절실한데, 그것이 바로 인공지능 윤리 규범이다. 인공지능 윤리와 관련한 논의에서 슈퍼인공지능 등 인공지능에 대한 지나친 기대로 인해 그 방향에서 다소 혼선이 있는 것도 사실이다. 인공지능은 먼 미래의 일이 아니라 현실이며, 그 대응책의 모색도 현실에 기초하여야 한다는 관점에서 보면 인공지능의 자율성에 주목하여 인공지능의 주체성이나 법인격부여 등을 논의하는 것은 시기상조라고 여겨진다.

인공지능 윤리 개념 논의는 인공지능으로 인한 부작용의 방지와 활용의 바람직한 방향의 모색에서 출발해야 한다. 그런데도 국내에서 한동안 인공지능이 어떠하여야 하는가 하는 인공지능 자체의 윤리에 대한 논의가 인공지능의 주체성 등의 논의와 함께 이어지다가 최근에 이르러 비로소 인공지능 자체의 윤리가 아닌, 인공지능 개발이 어떠한 방향으로 진행되어야 하는가 하는 개발하는 사람의 윤리가 진정한 인공지능 윤리라는 인식 전환이 일어나면서, 인공지능 윤리 논의는 제자리를 찾아가고 있다.

제1장 도입

인공지능이 도시기반시설을 비롯하여 법 집행, 금융, 헬스케어, 인도적 지원이나 심지어 남녀 간의 데이트 어플리케이션에까지 적용되면서, 인공지능은 머신러닝 알고리즘 그 자체로, 또는 로봇에 이식되기도 하면서 경제나 사회 복지 향상, 인권 신장 등에 기여하고 있다. 이러한 인공지능의 등장은 일자리 소멸이나 소득불평등의 심화 등 여러 가지 문제가 초래하면서 이에 대한 대책의 강구가 시급한 것도 사실이다. 또한 고위험 분야에 인공지능이 활용되면서 인공지능이 믿을 수 있고, 공정하며, 투명할 뿐만 아니라 안전하기를 바라는 요구 또한 높다. 이러한 요구를 어떻게 수용하고 해결할 것인지는 쉽지 않은 과제이다.

인공지능이 특정 영역에서 인간보다 우월한 성능을 보이면서, 인공지능이 합리성을 띤다거나, 일부 휴머노이드 로봇의 쇼에 현혹된 이들이 인공지능이 인간과 흡사하다며, 인공지능에게 법적 지위를 부여한다거나 도덕적 행위자로서의 윤리적 지위, 수범자로서의 의무이행 등을 거론하여 왔지만, 황당한 논의라고 일축하고 싶다. 국내에서 근래에 행하여진 수많은 법학적 논의들은 인공지능의 인격성에 대한 고찰을 비롯하여 인공지능을 책임주체로 볼 수 있는지, 인공지능 자체에 대하여 책임을 어떻게 지울 수 있을 것인지에 상당한 비중을 두어왔다. 그러나 이러한 논의는 현실의 인공지능과 무관한 논의이며, 그런 연구의 필요성이 전혀 없는 것은 아니나 연구의 중심이 되어서는 곤란하다. 인공지능과 관련한 특이점이나 슈퍼지능에 대하여 확신하고 있는 외국의 저명한 학자들도 그와 같은 논의와는 일정한 거리를 둔다.

인공지능과 관련하여 어떤 대응 프레임이 필요할까. 인공지능의 불확실성 및 예측 불가능성으로 인한 사회적 부작용을 최소화하고 개발자, 이용자 모두에게 이로운 방향으로의 활용을 위한 윤리 규범의 정립이 필요하고, 법적 장치에 의한 보장도 필요한 것은 분명하다(Cath, 2018). 이런 윤리 규범 등의 정립에 있어 사업자, 개발자, 이용자 등 참여주체뿐만 아니라 공적 주체가 함께 하는 공론화 과정을 거칠 필요가 있다. 그 과정에서 주도권은 사업자나 개발자, 이용자 등 민간이 가지되 공적 주체는 갈등 조정 등의 기능을 담당하여야 할 것이다. 기업이나 개발자는 이윤추구 목적에서 인공지능 시스템 개발을 하는 것이므로, 민간의 완전한 자율에 맡기는 것은 문제가 있다. 따라서 공적 주체가 참여하여 최소 기준을 제시하는 등 일정한 규제역할을 반드시 담당하여야 한다.

제2장 기술과 응용

제1절 서론

인공지능의 자율성(Autonomy), 예측곤란성(Unforeseeability)라는 특질은 인공지능과 관련한 윤리적 대응이나 정책 수립, 법적 규율에서 우선적으로 고려하여야 할 요소다. 인공지능 특히 최근의 머신러닝 시스템의 경우 비록 그 알고리즘이 잘 알려져 있다고 하더라도 모든 상황에서의 작동을 예측하긴 어렵다. 머신러닝 시스템의 작동은 데이터 세트의 패턴과 상관관계에 의하여 이루어지는데, 그 특성상 예측하기 어렵다.(Burri, 2016.).

인공지능의 자율성은 스스로 어떤 결정을 할지 선택하고, 그 결정을 실행할 수 있는 능력을 의미하는데(Gunderson & Gunderson, 2016), 인간의 직접적 지시 없이 기계가 획득하고 분석한 정보를 바탕으로 독립적으로 행동한다는 점 또한 자율성의 속성이다(Vladeck, 2014). 머신러닝의 특장점은 인간의 개입 없이 스스로 학습을 통해 향상된다는 점인데, 이러한 학습과정 때문에 생기는 문제도 없지 않다. 예를 들면 마이크로소프트가 개발한 챗봇 'Tay'는 공개하자마자 혐오발언 등을 쏟아내어 서비스를 중단할 수밖에 없었다. Tay가 사람들과의 대화 과정에서 오도된 학습을 하였기 때문인데, 이와 같이 학습에 사용된 데이터의 질에 따라 알고리즘이 문제행동을 일으키거나 차별이나 편향성의 문제점을 노정할 가능성은 아주 높다. 심지어 인공지능 알고리즘을 적용한 번역기는 성차별의 경향을 보이기도 한다.

제2절 인공지능으로 인한 사회경제적·윤리적 파급효과

인공지능 확산으로 인해 인류의 삶에 미치는 직 · 간접적 파급효과는 다양하다. 그 중 가장 흔히 회자되고 심각한 것은 소위 인공지능이 인간의 주요 직업을 대체함에 따른 일자리 상실의 우려이다. 변호사나 의사 같은 전문직조차 대체될 것이라는 전망이 난무하면서 불안감이 증폭된다. 분명 인공지능이 인간보다 더 효율적으로 수행해낼 수 있는 영역이 있다. 대체로 단순 · 반복적인 업무가 그럴 것이라고 진단한다. 이러한 인공지능에 의한 대체로 인하여 실직한 사람들이 무엇을 하며 살 것인지에 대한 다양한 대책도 논의되는 상황이다. Elon Musk는 인공지능으로 인해 실직한 사람들이 아무 일을 하지 않아도 급여가 주어질 것이라고 예측하였다. 그러나 이런 전망과 더불어 인공지능으로 인하여 새로운 일자리도 창출될 것이고, 인공지능 확산이 꼭 대량 실직으로 이어지는 것은 아니라는 전망도 나온다.

인공지능 로봇이나 알고리즘의 등장으로 인해 인간은 더 이상 실제 인간보다는 기계와 빈번히 접촉하고 교감할 것이며, 외부와 단절한 상태에서 오로지 기계와 소통하는 인간관계의 퇴화나 인간성 상실 또한 우려된다.

이런 문제보다 더 심각하고 현재화된 문제는 편향성, 차별성의 폐해다. 그동안 실전에 배치된 인공지능 알고리즘이 편향된 데이터를 학습한 결과 편향성을 보이고, 그로 인해 그 인공지능 알고리즘이 처리한 결과가 인종이나 성별에서 차별성을 보인다는 여러 가지 실례가 나오면서 인공지능의 편향성이나 차별성의 문제는 사회적으로 그 반향을 일으켰다. 대표적 사례가 범죄예측 시스템이 흑인에게 불리한 결과를 보인다는 것이다. 이런 상황 때문에 인간의 편향성이 인공지능에서 재생산된

다는 문제를 해결하여 차별적 결과 산출을 막는 것은 인공지능 윤리 이슈의 시급한 과제로 등장할 수밖에 없는 것이다. 또 다른 차원의 문제는 인공지능의 불완전성(Artificial stupidity)의 문제다. 인공지능이 제공된 학습 데이터로 제대로 학습을 하였다고 하더라도 그러한 훈련 과정이 실제 세계의 가능한 모든 경우를 대비하긴 어렵다. 그 결과 인간이라면 도저히 할 수 없는 어리석은 행동을 할 수 있다. 따라서 전적으로 인공지능에 의존하거나 지나친 힘을 부여해서는 아니 되고, 인공지능이 설계자의 의도대로 작동하는지를 끊임없이 확인하여야 안전성과 효율성을 보장할 수 있다. 이것이 보장되지 않으면 인공지능은 사악한 존재가 될 수 있다. 암을 치료하는 임무를 수행하기 위해서 암 환자를 죽이는 방법도 택할 가능성이 있는 것이 인공지능 알고리즘이다. 물론 환자를 살해함으로써 부여된 임무를 완수한다는 것은 설계자가 의도한 바가 아니다. 따라서 인공지능이 해결책을 찾을 때 임무의 전체적인 맥락을 제대로 이해할 수 있는 그런 알고리즘이 구현되는 것이 안전을 위해서도 필요하다.

〈표1〉 인공지능이 유발하는 위험 가능성 (출처: Leslie, 2019)

Potential Harms Caused by AI Systems
Bias and Discrimination
Denial of Individual Autonomy, Recourse, and Rights
Non-transparent, Unexplainable, or Unjustifiable Outcomes
Invasion of Privacy
Isolation and Disintegration of Social Connection
Unreliable, Unsafe, or Poor-Quality Outcomes

국 · 내외에서 인공지능 윤리 이슈와 관련하여 다양한 분야가 주목을 받고 있다. 그러나 거론되는 분야 중 상당한 부분이 인공지능 윤리 이슈와 무관한 것도 있다.

제3절 주요 인공지능 윤리 이슈

1. 자율주행차량

자율주행자동차는 제조분야의 윤리 이슈로 가장 먼저 떠오른다. 자율주행자동차는 인공지능의 상용화 과정에서 중요한 위치를 점하고 있다. 1986년에 이미 첫 선을 보인 이후 지속적인 투자대상이 되고 있고, 기존 자동차 제조업체는 물론이고 테슬라, 구글, 우버 등 첨단기술을 앞세운 업체들도 가세하고 있어 자동차산업은 인공지능을 중심으로 재편되리라는 예상이 파다하다. 반면 인공지능 알고리즘의 하자로 인한 사고로 사망자가 발생하는 등 그 안전성에 대한 우려도 크다.

우리나라도 2019. 4. 30. 자율주행자동차 상용화 촉진 및 지원에 관한 법률을 제정하여 자율주행자동차와 관련한 여러 가지 특례 등을 마련하는 등 자율주행자동차의 도입 · 확산 및 자율주행자동차의 상용화를 촉진하기 위한 정책을 시행하고 있다.

자율주행자동차가 인간 개입 없이 도로상황 등 외부 환경을 변수로 받아들여 스스로 의사결정을 한다는 차원에서 자율성을 중심화두에 놓기도 하지만, 이런 속성에 구애되어 도덕행위자 여부나 행위주체성 논의에 빠져서는 아니 된다. 인공지능의 알고리즘에 기초한 판단 및 행동을 모든 상황에 대해 예측할 수 없고 그로 인해 책임 규명이 다소 어려울 수도 있지만, 윤리적 대응이나 법적 규율이 불가능한 것은 아니다. 자율주행자동차를 도덕행위자 또는 행위주체자로 보고 자율주행자동차를 책임의주체로 삼는다면, 정작 책임을 져야할 개발자나 사용자 등의 책임을 면제하는 잘못된 논리로 비화될 가능성이 있으며, 이것은 현재의 인공지능 윤리 대응 노력에 반하는 것이라고 할 수 있다.

2. 자율 무기

자율 무기 개발도 가속화되고 있다. 강대국들은 자율무기 분야에서 선두주자가 되기 위해 각축을 벌이고 있다. 무장드론무기, 자동전투기, 전투로봇은 이미 실전배치가 되었다. 자율무기는 인간 전투원보다 훨씬 효율적이고, 강력하다. 이런 상황은 로봇 군대를 가진 나라와 그렇지 못한 나라 간의 비대칭 상황이 연출된다는 점에서 국가안보 차원의 문제로 직결된다.

이러한 자율무기 개발은 인류에게 직접적 위협이 될 수 있다는 점에서 이러한 무기는 금지되어야 한다는 주장이 대체로 힘을 얻고 있다. 그러나 자율무기의 대량살상무기로서의 성격은 인공지능 윤리와 직접 관련이 없다. 대량살상무기와 관련된 논의는 군축 등의 차원에서 다뤄져야지 인공지능 윤리에서 언급할 문제는 아니다. 오히려 자율무기의 속성 중 살상에 이르는 의사결정에 인간이 개입 하지 않는다는 점에 주목하여야 한다. 윤리적 관점에서 보면 기계적 판단에 의한 살상과 인간의 판단에 의한 살상은 전혀 다른 문제다. 문제는 사람들이 얼마나 많은 도덕적 책임을 기계에 아웃소싱 하느냐이다. 인간을 살상한다는 중요한 의사결정을 전적으로 기계에 맡기는 것은 윤리적 관점에서 도저히 허용될 수 없는 일이다. 유엔은 의사결정 과정에 사람들이 참석하지 않는 한 자율 무기의 사용을 금지하는 무기통제협정을 만들자고 제안했다(Morris, 2016). 그것을 인공지능 윤리의 관점에서 해석하면, 그와 같은 자율무기를 설계하여서는 아니 된다는 의미일 것이다. 인공지능을 이용한 무기 개발을 금지한다는 것은 기술의 관점에서 받아들이기 어렵다. 무기체계가 인공지능을 이용하여 보다 효율적이고 안전하게 될 수 있다면, 인공지능 알고리즘의 도입을 말릴 이유는 없다. 이런 측면에서 보면 인공지능 컨퍼런스에서 인공지능 무기 경쟁을 경고하는 성명서를 발표하거나 세계의 유력인사들이 인공지능 무기에 반대하는 성명서를 발표하는 것은 이상하다. 인공지능 무기 전부가 아니라, 살상과 같은 의사결정을 오로지 기계에 미루는 그와 같은 자율무기의 개발을 저지해야 한다.

3. 보건·의료분야의 인공지능

인공지능 알고리즘과 데이터의 활용으로 인해 보건의료 분야도 많은 변혁이 예상된다. 의료기술의 향상이나 의료서비스의 개선을 위해 인공지능 기술이 도입될 것이라는 점은 분명하다. 이와 같은 상황에서 보건 · 의료 분야에서 주된 인공지능 윤리 이슈는 데이터 활용에 있어서 환자의 개인정보 유출이나 재산권 침해 그리고 인공지능 분석 결과에 대한 책임 문제라고 할 것이다. 의료분야의 인공지능 알고리즘을 위한 학습 데이터는 기왕에 치료한 환자들의 개인정보이기도 하다. 또 인공지능 알고리즘의 진단이 잘못되거나, 오작동으로 인한 의료사고 발생 시의 책임 소재 문제도 대두될 가능성이 있다. 또한 환자들이 의사의 의견보다 인공지능의 진단을 선호하는 경향이 생기면서, 의료진이 인공지능 알고리즘의 판단에 전적으로 의존하는 의존증이 문제될 수도 있다. 의료인공지능의 실효성, 안전성의 평가 등은 의료분야에 국한된 문제는 아니라 할 것이다. 인공지능에게 태아를 살릴 것인가, 산모를 살릴 것인가 하는 결정을 맡겨서는 아니 된다.

법적 규율의 관점에서는 인공지능 윤리 이슈와 무관한 문제들이 생겨날 수도 있다. 인공지능을 이용한 헬스케어의 의료행위 해당 여부가 바로 그것인데, 그것은 윤리적 측면보다는 법 정책의 문제와 연관되어 있다.

4. 재범위험성 예측 프로그램

미국은 재범위험성을 범 집행의 중요한 잣대로 여기고, 이를 근거로 피의자나 피고인의 석방, 보석 등 여러 가지 법 집행을 행해왔다. 종전에는 법 집행 관계자의 경험 등을 바탕으로 재범의 위험성을 예측하여 왔으나 점차 통계적 기법을 활용하여 재범의 위험성을 과학적으로 예측하는 방향으로 정책 전환이 이루어지고, 근래에서 이르러 인공지능 알고리즘으로 재범의 위험성을 예측하기에 이르렀다. 미국에서 널리 사용되는 재범 위험성 예측 알고리즘은 COMPAS(Correctional Offender Management Profiling for Alternative Sanctions)인데, 이러한 COMPAS의 재범예측 결과의 신뢰성에 대하여 ProPublica는 의문을 제기하고 나섰다(Kirchner & Angwin & Mattu & Larson, 2016). 흑인과 백인 간의 예측률에서 흑인에게 불리한 결과가 나왔기 때문이다.

〈표2〉 재범예측 알고리즘의 인종 별 판단 (출처: ProPublica)

	white	African American
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

이러한 편향의 문제 발생에 대하여도 다양한 원인 분석이 이루어졌다. 주로 편향된 데이터를 학습한 알고리즘의 문제라는 것이다. 인공지능이 편향된 데이터에 의하여 학습되지 않도록 하는 윤리적 이슈와 직결된 문제인데, 데이터의 확보 과정에서 나름대로 원칙이 될 수 있는 가이드라인이 제공되어야 한다.

5. 로보어드바이저

로보어드바이저 서비스는 금융회사와 소비자 간의 정보비대칭 문제를 심화할 수 있고, 시스템 오류로 인한 시장의 혼란 상황도 이미 발생한 바 있고, 향후로도 우려된다. 그러나 이와 같은 문제를 인공지능 윤리 이슈로 포장하는 것은 문제가 있다. 이것은 기존 시스템에도 내재된 위험성이며, 인간의 실수에 의해서도 발생되는, 인공지능 알고리즘이 도입됨으로 인해서 새로 생긴 문제는 아니다. 따라서 그 해결책도 인공지능 윤리 차원에서 모색할 성격이 아니다. 이윤추구가 주목적인 금융 분야 알고리즘 구현에서 윤리적 차원의 행동규칙을 프로그래밍 해 넣어야 한다는 것은 무리가 있다. 문제가 발생해도 대부분 고객보호를 위한 일반적인 법적 규율의 대상이 될 뿐이다. 법적 규율의 대상이라고 하여 반드시 윤리와 연관시킬 필요는 없다. 로보어드바이저에 사용된 알고리즘은 적어도 충분한 전문성을 갖춘 투자전문가가 개발하고, 감시에도 참여해야 된다는 것이 윤리적 설계인가? 알고리즘 가정(assumption)은 일반적으로 인정된 투자이론에 기초해야 하고, 알고리즘 가정에 대한 설명서를 알기 쉬운 언어로 작성하여 고객이 언제라도 볼 수 있도록 공시하여야 한다는 원칙이 금융 분야 인공지능 알고리즘의 윤리로 적절한가? 그런데 그러한 원칙을 인공지능 윤리로 거론하고 있는 것이 현실이다.

제 3 장 인공지능 윤리 정립 방안 논의

제1절 왜 인공지능 윤리인가

인공지능의 가능성 중 예측기능의 중요성은 크다. 이러한 예측기능은 기존의 컴퓨터 알고리즘으로는 구현하기 어려웠던 임무를 가능케 한다. 이러한 예측기능으로 인해 인공지능은 보다 지능적이라고 여겨지는 영역에서 인간 대신 판단하는 일을 수행하고 있고, 인간의 개입 없이 오로지 인공지능의 판단과 예측에 의하여 주요한 의사결정이 이루어지고 있는 실정이다.

인공지능이 사회에 미치는 영향력이 점점 더 커지고 있는 상황에서 인공지능의 활용은 긍정적인 결과로 이어지는 것이 대부분이지만, 그로 인한 부작용도 적지 않다. 특히 인공지능 알고리즘의 불투명성과 복잡성은 인공지능 활용에 있어서 넘어야 할 큰 산이다. 인공지능의 활용은 사적 영역뿐만 아니라, 공적 영역에서도 뚜렷하다. 인공지능 알고리즘의 복잡성과 편향성이 커지면서 윤리적 우려 또한 커지고 있다.

현재 문제되는 윤리적 쟁점은 편향성, 불공정성, 안전성 결여, 불투명성 및 책임소재 불분명 등에 관한 것인데, 사적 영역에서의 활용으로 인한 폐해도 적지 않지만, 공적 영역에서의 인공지능 활용으로 인한 폐해는 그 파급효과가 크다. 예를 들면 법 집행과 같은 공공영역에서의 인공지능이 인종차별이나 성차별과 같은 결과를 보일 때에는 그로 인한 사회적 여파는 심각하다. 인종이나 성차별과 같은 문제는 중대한 정치 · 사회적 문제이기 때문이다.

제2절 인공지능이 윤리 또는 법적 책임의 주체인가?

인공지능에 대한 논의가 자칫 엉뚱한 방향으로 흘러서는 곤란하다. 인공지능과 관련하여 가장 빈번하게 거론되는 것이 자율주행자동차와 관련한 법적 윤리적 쟁점이고, 나아가 인공지능에게 어느 정도의 자율성을 부여하고 도덕적 결정권을 부여하는 문제가 인공지능 윤리인양 착각하는 현상도 나타난다. 나아가 자연스럽게 인공지능이 야기한 해악에 대하여 인공지능 자체에 대하여 법적 책임을 물을 수 있는가 하는 논의로 이어지기까지 한다.

알파고의 등장이 인공지능에 대한 인식에 대변환을 가져온 것은 분명하지만, 너무 지나치게 그 의미를 침소봉대하는 면이 없지 않다. 강화학습 알고리즘의 일종에게 인간의 직접적 개입이나 간섭 없이 일종의 직권 위임에 따라 스스로의 선택이나 판단을 통해 자율적으로 행동하는 인공지능 행위자 지위를 부여하려는 일부의 견해는 너무 터무니없다. 비록 바둑 등 특정 영역에서 인공지능이 인간을 패퇴시켰지만, 인공지능의 능력이 인간에 비해 현저히 떨어진다는 점을 부정하는 인공지능학자들은 없다. 뿐만 아니라, 인공지능에게 도덕적 상태가 없다는 점을 부인하지도 아니한다(Bostrom & Yudkowsky, 2011).

현재의 인공지능의 핵심은 머신러닝인데, 이는 주어진 데이터로 입력 값과 출력 값 사이의 상관관계를 통해 일정한 패턴을 파악하는 학습 기능이 핵심이며, 이를 통해 예측 · 분석 기능을 수행하는 것에 불과하다. 어떻게 보면 단순한 학습능력을 갖춘 알고리즘을 너무 지나치게 과대평가하는 것은 문제가 아닐 수 없다.

제3절 인공지능 윤리나 책임논의의 방향

인공지능의 특성은 법제도적 차원의 접근에서도 다른 특성을 갖는다. 인공지능의 엄청난 파급효과를 감안하면, 통상의 법에 의한 규율방식인 사후적 구제책보다는 사전적 예방이 강조되기도 한다. 이런 사전적 예방에 의한 규율이 강조되다 보면 그러한 규율로 인하여 개발이 위축될 수도 있다는 우려가 없을 수는 없는데, 그런 면에서 법적 규율 이전의 윤리적 차원의 해결이 더 강조되기도 한다. 그러나 인공지능의 윤리라는 것은 개발 단계에서 지켜야 할 일종의 수범규칙으로 개발윤리라고 해도 과언이 아니고, 그러한 윤리적 문제는 개발 단계에서 해결하여야 한다. 따라서 통상적인 주체성과 연관된 인간의 윤리와 인공지능 윤리는 본질적인 차이를 보인다. 인공지능 윤리는 인공지능 기술이 어떤 식으로 사회에 수용되고, 사회로부터 어떻게 인정받느냐의 문제이며, 그러기 위해서 인공지능 기술이 어떤 특질을 지녀야 하는가 하는 문제이기도 하다. 결국 인공지능 윤리는 인공지능 자체가 아닌, 우리 사회가 인공지능에 요구하는 인공지능 개발의 조건으로 개발자 등 참여주체들이 지켜야 할 규범이라고 봐야한다. 따라서 인공지능 윤리는 결국 사회의 가치를 수용하고, 사회 구성원들의 다양한 요구와 의사를 반영할 수 있는 것이 될 수밖에 없다.

그러나 실제 이러한 윤리를 규정하고, 그에 따른 윤리적 대응을 하는 것은 쉽지 않다. 근본적 가치나 윤리 프레임은 너무 복잡해서 인공지능의 연역적 결정 시스템에 공식화하여 넣기가 어렵다((Leikas & Koivisto & Gotcheva, 2019). 윤리적 결정을 인공지능에서 구현한다는 것은 다분히 맥락 의존적인데, 전통적 접근방식으로는 어렵다.

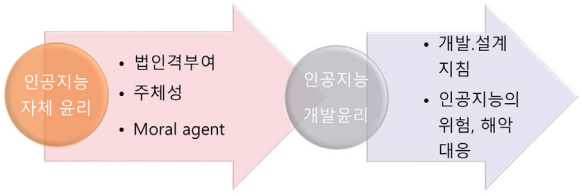
인공지능 윤리규범의 구상에 있어 우선적으로 고려하여야 할 것은 인공지능 자체에 대한 이해가 선행되어야 한다는 점이다. 인공지능 윤리를 논의하는 과정에서 인공지능 본질을 제대로 이해하지 못한 채 피상적으로 접근하는 바람에 얼토당토 않는 해법이 제시되기도 한다.

인공지능은 결코 인간을 그대로 묘사하는 식으로 개발하지 않는다. 인간의 방식과는 다르지만, 그 결과에 있어서 동일하면 족하다는 접근법이 특징이다. 결코 인간처럼 사고하지 않고, 임무를 수행하는 방식도 인간과는 다르다. 로봇 개발에서 이해하기 어려운 것이 이족보행 로봇에 대한 집착이다. 인간처럼 두 발로 걷는 로봇을 만들겠다는 의욕 때문이겠지만, 사족보행 로봇에 비해 매우 비효율적이고, 개발의 난이도도 높다. 이족으로 균형을 잡는 방법을 찾아내는 것 자체가 큰 난제이다. 그러나 로봇이 굳이 인간처럼 이족 보행할 이유가 없고, 바퀴가 아닌 다리로 이동할 수 있는 기능만 주어지면 그 발이 몇 개가 되어도 된다.

심지어 인공지능의 자율성에 주목하면서, 인공지능에게 지나치게 명료한 임무를 부여해서는 아니 된다는 견해도 있다. 인공지능에게 인간의 고통을 없애라는 구체적 임무를 부여할 경우, 그 고통을 없애는 방편으로 인간을 없애려고 할 수 있다는 예를 들기도 한다. 그러나 이런 문제는 명료한 임무를 부여해서라기보다는 인공지능이 임무 자체를 이해하지 못한 데서 나오는 문제다. 뿐만 아니라 그런 알고리즘 구현 과정에서 목적 달성 여부에만 신경을 쓰고, 그 수단의 정당성에 대한 고려가 없었기 때문이기도 하다는 점을 간과해서는 아니 된다.

컴퓨터의 성능 향상과 머신러닝에 사용될 수 있는 빅 데이터의 등장으로 인해 인공지능의 성능이 비약적으로 발전하고, 그 역할도 증대되었다는 점은 부인할 수 없다. 특히 사회적으로 큰 영향을 줄 수 있는 부문의 의사결정에 인공지능이 활용되면서 인공지능의 사회적 영향력에 따른 윤리나 책임에 대한 논의는 더 이상 늦출 수 없게 되었다. 인공지능이 바람직한 결과만 도출하지 않고 여러 가지 폐해를 가져오면서 소위 인공지능 또는 로봇 윤리(Robot Ethics)에 대한 논의의 필요성이 커졌다. 최근 자율주행자동차 운행 과정에서 사망사고까지 발생하면서 인공지능의 의사결정 과정을 설명해줄 수 있어야 한다는 요구도 나오고 있는 실정이다. 이는 투명성의 요구를 넘어 설명 가능한 인공지능의 요구라고 할 수 있다. 이와 같은 요구는 인공지능의 현실과 직결된 문제로 그 해결은 인공지능

의 성패와 직결될 문제다. 논의도 그런 부분에 집중되어야 한다. 문제는 인공지능 윤리와 관련된 국내 기존의 연구가 그런 방향으로만 흐르지 않았다는 것이다. 본래 윤리와 책임이 '주체성'과 연관을 맺을 수밖에 없다는 점에서 인공지능이나 로봇이 주체적인 존재가 될 수 있는 지를 논의의 시초로 삼았다. 그러다보니 자연 인간의 윤리를 인공지능에 덧입히려는 시도를 하게 되고, 인공지능의 윤리를 운위하기 위해서는 인공지능이 인간과 마찬가지로 주체성을 지난 존재라야 한다는 생각을 근저(根底)에 깔면서 한결같이 인공지능을 인간과 같이 인격체로 취급할 수 있느냐를 비롯하여 여러 가지 선행 문제에 집착할 수밖에 없었다. 로봇을 주체성을 가진 인격체로서 취급하는 것을 전제로 전재되는 인공지능 윤리는 현 단계에서 너무 비현실적이다. 그러나 최근 국내의 인공지능 윤리와 관련된 논의에 태도변화가 이루어지기 시작했다. 인공지능 윤리의 담론이 점차 인공지능 자체가 아닌 개발자의 윤리로 변화하기 시작했다. 더 이상 인공지능 자체의 윤리가 아닌 인공지능 알고리즘을 개발하고, 운용하는 사람에 대한 통제 원리를 모색하는 식으로 논의가 바뀌기 시작했다.



〈그림1〉 인공지능 윤리 담론의 진행 방향

제4절 인공지능이 아닌, 인공지능 개발의 윤리

인공지능이 어디에나 널리 이용되는 편재성遍在性)이 뚜렷해지면서 현실에서 작동하고 있는 인공지능이 어떤 식으로 작동되어야 하는가를 명확히 정하고 인공지능 작동으로 인하여 인간 사회에 이로움을 주고, 해악을 주어서는 아니 되며, 인공지능의 작동으로 인간사회가 위협받는 일이 없어야 한다는 점을 공고히 할 필요가 있다. 이런 역할을 하는 것이 바로 인공지능 윤리 규범이다. 인공지능 자체가 아니라, 인공지능을 개발하고 운영하는 규범으로서의 윤리에 주목하여야 한다.

인공지능 윤리는 인공지능 기술의 개발과 운용에 있어 무엇이 옳고 그른지에 대한 도덕적 지침으로 받아들일 수 있는 가치, 원칙, 기술 등으로 구성되는 세트로 볼 수 있다(Leslie, 2019). 다만 이러한 인공지능 윤리는 자율성이 특징으로 강제력이 부여되는 법규범과는 구분된다. 법 규범의 경우 구속성을 가질 뿐만 아니라, 통제 대상은 입법을 통해 결정된다. 반면 윤리규범의 경우 사회적 합의의 산물로 기존의 윤리는 대부분 자연스럽게 형성되고, 사회구성원이 이의 없이 받아들이면서 구성원 누구도 윤리 규범 자체에 대하여 딱히 의문을 가지지 않는다. 반면 인공지능과 같은 인위적 존재의 경우, 이와 같은 자연스러운 윤리 관념의 형성이 있을 수 없고, 충분한 운용 경험이 축적되지 않으면 윤리 관념의 생성도 어렵다. 법적 규율과 윤리적 규율의 경계는 분명하지만, 법 규범 자체가 윤리 규범과 전혀 별개의 것이 아니라 어느 정도 상관관계가 있다고 하면 인공지능을 대상으로 한 법 규범의 제정 또한 한계가 있을 수밖에 없다. 인공지능의 행위로 인한 책임의 바탕에는 윤리 규범 위반이 깔려있어야 하는데, 인공지능 윤리 규범의 정립이 명확하지 않은 단계에서는 책임 문제도 불명에 빠질 가능성이 높다. 따라서 인공지능 윤리 규범이 어떠하여야 하는지에 대한 규명이 선행되어야 한다.

인공지능 알고리즘은 설계자의 요구에 따라 최적화된 결과를 도출한다. 따라서 인공지능에 사용되는 데이터보다는 개발자가 그런 알고리즘에게 부여한 임무 요구가 알고리즘의 성능을 좌우한다. 따라서 인공지능 윤리에 포함될 가치나 원칙, 기술 등은 윤리적이고, 공정하며, 안전한 인공지능 어플리케이션을 생산하는데 필요한 기본적 책무를 설명하는 것이어야 한다. 또한 인공지능 시스템의 오용이나 남용, 설계 오류, 의도치 않는 부작용 등으로 인해 초래될 수 있는 개인이나 사회적 손실에 대한 대응책의 전제라는 점을 염두에 두고 규정되어야 한다.

제 4 장 인공지능 윤리 이슈의 정책적 대응

제1절 인공지능 윤리 대응 지침 개관

도덕성은 사회 수준에 따라 문화마다 다르다. 그것은 경향성을 띨 뿐 아니라 기술적 진보에 따라 끊임없이 변한다. 마찬가지로 윤리적 문제는 인공지능이 생성한 새로운 범주에 따라 점점 더 복잡해지고 있다. 옳고 그름의 이분법적 견해로는 합의가 이루어지지 않을 수 있다. 이런 문제에 대한 해결책은 그러한 윤리의 복잡한 본질을 이해하는 데 있다. 알고리즘의 목표는 가치와 윤리적 행위 측면에서 사람들의 목적에 부합하여야 한다. 그 가치와 도덕은 프로그래밍 수준에서 구현하든, 인공지능 스스로 학습과 관찰에 의해 획득하든 확실해야 한다 (Pavaloiu & Kose, 2017).

또 다른 해결책은 사람으로 하여금 책임을 지게 하는 방안이다. 인공지능의 기능적 오류이든 부주의로 초래된 것이든 오용이 되었든 인공지능이 초래한 잘못된 결과에 대하여 관여한 사람들에게 책임을 묻는 것이다. 인공지능 로봇 무기가 잘못 사용되었을 때도 인공지능 로봇 무기가 아닌, 그것을 개발하고, 실전에 배치한 사람들을 비난하고, 궁극적 책임을 묻는 형태의 윤리 규범으로 설계할 필요가 있다.

인공지능 윤리 지침의 설정에 있어 가장 중요한 것은 인간의 가치가 반영되어야 한다는 점이다. 물론 가치라는 요소는 시대와 사회에 따라 다양하며, 어떤 사회의 구성원 사이에서 합일의 가치관을 도출하는 것은 쉽지 않지만, 그래도 보편적인 가치라고 여겨지는 어떤 지점에서 최소한 사회구성원들이라면 의당 받아들여야 하는 'ought to be'가 윤리다. 인공지능으로 인해 야기되는 어떤 문제는 옳고 그름의 도덕적 잣대에 어긋날 수도 있고, 당위성이란 잣대를 충족할 수 없을 수도 있지만, 인공지능 윤리의 출발점은 사회 구성원들이 인식하는 당위성이 되어야 한다.

인공지능 윤리 규범을 정립하면 인공지능 개발 단계에서부터 이런 윤리 규범을 설계에 반영하여야 하며, 인공지능의 사용 승인 단계에서 이러한 윤리 규범 준수 여부를 평가하여 사용 승인 여부를 결정하는 잣대로 삼아야 한다.

인공지능이 가진 잠재적 위험은 실제 사용 과정을 통해 현실화되므로, 사전에 이를 미리 파악하고 예방한다는 것은 어렵다. 알고리즘 감사 등을 통한 사전 규율이 한계를 보이는 것도 이 때문이다. 따라서 윤리 규범의 준수라는 표지를 가지고 위험 진단을 시행하는 것이 현실적인 방안일 수 있다. 윤리 규범 준수에 있어 구성원들의 자율이 중요하다. 그러나 그런 규범이 제대로 준수되지 않아서 현실적인 피해가 발생하였을 때 자율적인 피해구제 노력은 한계를 지니게 되며, 그때부터는 윤리 규범을 넘어선 법적 대응이 필요하다. 이러한 법적 대응은 사용 승인이나 사용 과정의 감시와 같은 규율과는 다른 차원의 문제인 피해 구제책의 실현이라고 할 수 있다.

인공지능 윤리를 이야기하면서 대체로 인공지능 시스템은 인류와 지구를 위해 개발되어야 한다는 점과 인공지능의 혜택을 인류 모두가 공유하여야 한다는 점을 공통적으로 내세운다. 지극히 옳은 이야기이지만, 이는 도덕적 명제이지 윤리가 아니다. 인간의 윤리로도 쉽지 않은 이와 같은 허구적 당위성이 과연 인공지능 윤리로 가당키나 한 것인가? 인공지능 개발이나 사용에 있어 마땅히 지켜야 할 심리적 규범으로서의 윤리로는 어울리지 않는다. 인공지능 개발이 이윤추구를 목적으로 하는 거대 기업에서 주로 이루어지고 있는데, 이들에게 이와 같은 명제를 인공지능 개발 윤리라는 명분으로 강요할 수 있을까? 오히려 인공지능의 운용과정 전반에서 인간의 존엄성이나 완전성 그리고, 자유를 침해해서는 아니 되고, 사적영역(privacy)을 보호하며 문화와 성적 다양성을 존중하고 인권 침해가 되지 않도록 노력하라고 명령하는 것이 보다 더 명확한 인공지능 윤리가 아닐까? 인공지능 시스템에만 의존하지 않고, 인간이 통제권을 가져야 하는 알고리즘, 성(性)이나 인종, 성적 취향, 연령에 따라 인간을 차별하지 않는 알고리즘을 구현하기

위해서 개발 단계부터 사용에 이르기까지 모든 주의를 기울여야 한다는 것이 윤리 규범으로 맞지 않은가? 이와 같은 것이 인공지능 윤리의 가이드라인이 되어야 하는 것은 현재까지 인공지능으로 인하여 초래되거나 초래될 수 있는 해악과 무관하지 않다.

윤리는 사회구성원들의 경험을 바탕으로 한 인식에서 비롯된다. 어떤 행위가 사회에 해악을 가져온다면 이를 금지하여야 한다는 하는 식의 합의가 윤리규범이다. 인공지능이 실용화된 영역에서의 실제 경험이 막연한 관념보다 인공지능 윤리 규범을 형성하는 주요한 토대가 된다.

제2절 인공지능 윤리 대응의 구체적 방안

1. 가치 반영 설계

인공지능 개발 과정에서의 설계는 가치를 반영하여야 한다. 그 가치는 도덕적 가치(moral values)가 아닌 윤리적 가치(ethical values)를 의미한다. 도덕적 잣대를 실천 윤리 규범에 주입하여서는 아니 된다. 윤리 규범은 개개인의 판단에 따라 달라지는 도덕적 가치가 아니라, 사회 구성원들 간에 공통적으로 형성된 윤리적 가치에 기초하여야 한다. 인공지능에 대한 사회 구성원들의 경험이 반영되어야 하는 실천의 영역이며, 인공지능의 사용 경험이 축적되면 이러한 윤리 규범과 그 바탕의 윤리적 가치도 변할 것이다.

2. 공정성 확보

가. 데이터 공정성(Data fairness)

인공지능 알고리즘 특히 머신러닝 알고리즘의 경우, 훈련 데이터(training data)에서 통계적 패턴을 추출하도록 설계된 것인데, 만일 훈련 데이터가 소수집단에 대한 기존 사회의 편견을 반영할 경우 그 알고리즘은 그러한 편향성을 그대로 내포하게 될 가능성이 높다. 이와 같이 인공지능의 성패는 좋은 데이터의 확보에 달려있다고 해도 과언이 아니다. 제대로 된 데이터를 확보하여 가공하고 관리하여야 알고리즘 소기의 성과를 낼 수 있는데, 인공지능 알고리즘이 편향성이나 차별성을 보이지 않기 위해서는 편향성 등에 오염되지 아니한 데이터의 확보와 선별, 정제 등의 관리가 선행되어야 한다. 데이터가 편향되거나 오염 또는 왜곡되었다면 차별적 결과나 오류의 결과를 방지할 방도가 없다고 해도 무리가 아니다. 따라서 데이터를 다루는데 있어 공정성을 유지하는데 각별한 주의를 기울여야 한다(Leslie, D. 2019). 알고리즘의 편향성 문제를 어쩔 수 없는 내재적 결함으로 여기고 인공지능 윤리의 구상에도 이런 점이 반영되어야 한다는 데 초점이 모아지기도 했지만, 공정성을 해치는 이와 같은 편향성 등의 문제는 내재적 결함이라기보다는 데이터 공정성과 관련된 문제에서 기인될 가능성이 높다. 사실 데이터에 편견이 반영되었다는 것은 일종의 변명이다. 데이터 수집 후 분석을 통하여 그런 데이터의 문제점을 발굴하고 가려냄으로써 편향되지 않은 데이터를 머신러닝 알고리즘에 제공하여야 한다. 따라서 편향성이나 차별성이 보이는 결과 자체만을 비난할 것이 아니라, 그와 같은 결과를 초래한 이유에 대한 진지한 성찰과 분석이 따라야 한다.

나. 공정성 인식설계(Design Fairness)

개발자 등 참여주체는 인공지능 시스템 구축의 모든 단계에 관여하고 있다. 공정성 인식 설계는 이러한 참여주체들이 인공지능 프로젝트 작업흐름 전반에 걸쳐 편견이 개입되어 차별적인 영향을 미치지 않도록 주의의무를 다하는 것을 의미한다. 인종과 같은 속성을 분류에 명시적으로 사용하지 않아도 알고리즘이 편향된 결과를 보일 수 있다. 따라서 설계 과정에서 차

별적 결과로 이어질 수 있는 숨은 알고리즘의 구조적 편향성을 발견할 수 있는 방안이 강구되어야 한다. 데이터 전처리과정에서부터 머신러닝에 변수로 사용될 특성(feature) 결정과 모델 구축, 알고리즘의 구조적 분석 등에 이르기까지 모든 개발 과정에서 이와 같은 공정성 인식 설계 관념이 투영되어야 한다.

다. 공정 결과 도출(Outcome fairness)

차별적 위험에 대한 최소한의 보호책의 일환으로, 개발하고 있는 인공지능 시스템의 영향과 결과의 공정성을 정의하고 측정하는 방법을 고려하여야 한다. 인공지능 알고리즘은 객관적일 것이라고 착각하지만 그렇지 않은 경우가 많다. 그러나 이러한 편향성이나 차별성은 특정한 의도를 가지고 프로그래밍 해야만 나타나는 것은 아니다. 이런 편향성이나 차별적 결과는 인공지능 알고리즘의 본질적 속성일 뿐 프로그래머가 특정한 요인이나 변수에 가중치를 부당하게 부여하는 등 의도적 조작을 했기 때문에 나타나는 것은 아니다(Barocas & Selbst, 2016). 따라서 알고리즘 설계 단계에서 인공지능이 공정한 결과를 도출하도록 유념하여야 한다.

3. 공공성 확보

인공지능 기술로 인한 혜택은 인류 모두가 공유해야 한다는 것은 구두선(口頭禪)에 그칠 수 있다는 점에서 공공성 확보를 윤리규범으로 이야기하는 것이 옳은가? 이것이 정당화되는 지점은 인공지능 개발의 위험성이다. 즉 인공지능은 상당한 위험 속에 이루어지고 있고, 그 위험은 사회공동체가 감수해야 하는 것인 만큼 그로 인한 이익도 사회공동체에게 분배되어야 한다는 점에서 수긍이 간다. 인공지능으로 인한 위험과 편익에 대한 평가도 공적 영역에서 이루어져야 한다. 편익을 넘어서는 위험이 있다는 평가가 있으면 그런 인공지능 개발은 허용되어서는 아니 된다. 또 다른 문제는 인공지능 도입으로 인한 일자리 상실의 문제다. 인공지능 도입으로 인류 대부분이 직업을 잃은 상황이 되면, 인류 공존의 차원에서 이익의 분배 문제를 재고할 가능성이 있다. 어떻든 인공지능 알고리즘의 도입과 수용에서 사회적 합의가 중요하고, 그 과정에서 합의를 도출하기 위한 사회적 협의체도 필요하다.

4. 투명성·설명가능성 담보

다양한 형태의 자동화와 복잡성이 특징인 인공지능의 기계학습 모델은 차별적결과로 인해 필연적으로 불투명성과 책임성 문제를 야기하다보니 인공지능 맥락에서의 엄격한 투명성, 책임성이 윤리적 대응 방안으로 등장한다. 인공지능 개발 기업의 영업비밀 보호 필요성 때문에 투명성 확보가 어려울 수 있을 뿐만 아니라, 설사 그 알고리즘 코드를 공개하였다고 하더라도 의사결정 시스템에서 어떤 알고리즘이 실제로 사용되었는지 또는 처음에 프로그래밍 된 대로 작동했는지 여부를 알 수가 없다. 따라서 투명성과 그에 기초한 책임성을 따져보기 위해서는 인공지능 알고리즘의 설명 가능성(explainability) 문제가 해결되어야 한다. 설명가능성은 알고리즘의 입력과 출력, 그리고 그 근저의 여러 가정 사이의 인터페이스의 실제 조건과 데이터 세트가 어떻게 훈련되고 구현되었는지에 관한 문제를 해결하기 위한 해법으로 등장한 것이다. 이러한 설명가능성의 문제가 해결되면 인공지능 알고리즘에 대한 감사가 용이해지고, 사용자 또한 알고리즘 의사결정이 산출하는 결과물에 대한 적절한 정보를 얻을 수 있을 것이다. 또한 이를 통해 인공지능의 차별적 결과 등 인권 침해 문제에 대한 적절한 통제를 할 수도 있다. 그러나 그동안 기라성 같은 연구진들이 꽤 오랫동안 설명가능성 문제에 매달렸지만, 아직 소기의 성과를 거두고 있지 못한 실정이다.

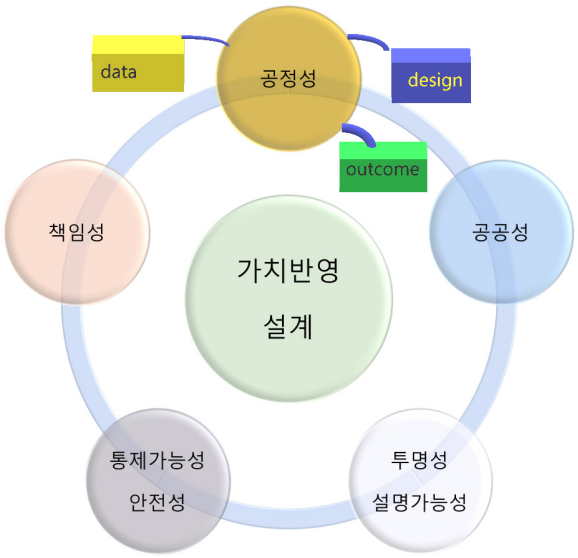
5. 통제가능성·안전성 확보

인공지능이 가진 자율성, 불투명성이 부각되면서 통제가능성이란 문제도 개발 윤리의 주요한 위치를 점하게 되었다. 자율성이나 불투명성은 윤리 문제를 넘어 법적 책임 규명 과정에서 큰 장애가 된다. 따라서 현실적 어려움에도 불구하고, 가능하면 자율성 · 불투명성을 극복하고, 인공지능을 통제할 수 있는 방안이 강구하여야 한다.

인공지능의 안전성 또한 중요한 문제다. 인공지능의 안전성은 기술적으로 지속가능하여야 하고, 정확성이 담보되어야 하며, 신뢰할 수 있어야 가능하다. 그러나 인공지능 시스템은 불확실하고 변화무쌍한 실세계에서 인간의 예상과는 다르게 작동될 위험이 있으므로, 안전하고 신뢰할 수 있는 인공지능 시스템을 구축하는 과제는 만만치 않다. 특히 인공지능 로봇의 오작동의 경우 인명손상으로 이어질 가능성이 높은데, 이런 경우 Asimov의 법칙과 같은 규칙기반 접근은 해결책이 되지 못한다(Sparkes, 2006). 로봇에게 결정을 맡기는 구조가 아니라 인명과 관련된 로봇의 행동을 인간의 통제 하에 두는 구조가 되어야 한다. 예를 들면 군사용 로봇이 스스로의 판단에 따라 인명을 살상하도록 해놓고, 인간이 개입하여 그런 살상을 막거나 중지시키는 방안을 모색하는 것은 본말이 전도된 잘못이다.

6. 책임성 확보

인공지능의 책임은 결국 인공지능으로 인한 문제가 생겼을 때 누구를 비난해야 하는지, 어떤 부분을 비난해야 하는지 하는 문제로 직결된다. 인공지능의 자율성, 불투명성에 주목한 이들은 개발자나 사용자도 예측하지 못한 결과에 대하여 개발자 등에게 책임을 묻기 어려운 문제의 해결방안으로 인공지능 자체에 대한 책임 부여 방안에 이르기까지 다양한 해법을 제시하고 있다. 그 어떤 것도 현재의 법 제도 하에서 받아들이기 어려운 부분이며, 향후 이런 문제를 수용할 수 있는 법 제도의 변혁이 필요한 영역이기도 하지만, 인공지능 윤리 차원에서는 개발 과정에서의 윤리 규범 위반에 대한 책임을 지는 문제로 환원해야 한다. 문제의 결과가 개발 과정에서 어떤 윤리규범을 제대로 준수하지 아니하여서 발생하였다는 전제 하에 역으로 개발자 등이 이런 윤리규범을 제대로 지켰는데도 불구하고 그런 결과가 발생하였다는 설명책임을 지도록 하고 그에 따라 개발자 등의 책임을 규명하는 것이 해법일 수 있다.



〈그림2〉 인공지능 윤리 대응 프레임

제3절 법제도적 대응

1. 법제도적 규율 일반

앞서 논의한 윤리규범의 이행은 어디까지나 개발자 등의 자율에 맡겨져 있다. 수범자의 자율성이 보장되고 그 적용에 있어 유연성이 담보되는 윤리 규범 차원의 대응은 한계가 있을 수 밖에 없다. 각 산업 분야별로 인공지능 기술을 적용할 때 이러한 윤리 규범을 구체적으로 관철시키기 위한 정책 대응이 따라야 한다. 인공지능의 윤리 규범 자체가 인공지능 자체의 윤리가 아닌 인공지능 개발 · 사용자의 개발윤리라는 결론이 타당하다면 인공지능에 따른 법적 규율도 그런 관점에서 개발 행위 자체를 대상으로 이루어져야 한다. 따라서 인공지능 알고리즘이나 로봇에 대하여 책임을 지우려는 시도는 타당성이 전혀 없다. 인공지능을 책임의 주체로 볼 것인가 하는 논의의 비현실성을 차치하고라도 인공지능의 오류 결과에 대하여 인간에게 책임을 묻는다는 원칙을 흐릴 수 있기 때문이다. 이런 엉뚱한 집착으로 인하여 정작 책임을 져야 할 인간에게 책임을 물을 수 없는 규율의 공백을 초래할 우려도 없지 않다. 인공지능 자체를 비난해서는 아니 된다. 그 뒤에 숨은 인간은 비난해야 한다.(Hofheinz, 2018). 즉 개발에 관여한 인간에게 그 개발 과정에서 요구되는 개발 윤리를 준수하지 못한데 대한 책임을 물어야 한다.

흔히 인공지능에서 책임 문제가 어려운 이유로 인공지능의 불투명성, 자율성 등이 거론된다. 특히 머신러닝과 같은 경우, 개발자조차 숨은 연산과정을 추적하기가 쉽지 않다. 이런 문제점을 들어 개발자에게 책임을 묻는 데 난관이 있다고 들 한다. 그렇다고 해서 책임 논의를 미룰 수는 없다. 인공지능이 점차 인간 생활에 간여하는 비중이 높아지고 있는 상황에서 명확한 책임의 가이드라인을 정할 필요가 있다. 우리 인간은 신기술의 책임 문제에 있어 결코 선제적이질 못하였다. 빈번하게 일어나는 인류의 비극은 제대로 사용할 방법을 배우기 전에 어떤 물건을 발명한다는 것이다. 원자폭탄의 경우가 그렇다. 그로 인한 비극적 결과가 발생한 후에야 원자폭탄을 규제하기 위한 전 세계적 기구가 마련되고 규율을 논의하기 시작했다.

인공지능이 인류 모두의 궁극적 이익을 사용될 것인지, 아니면 인류 갈등의 골을 깊게 하고 악화시킬 것인지는 인공지능 윤리가 수립되고 그에 따라 개발이 이루어지느냐에 달려있는 것이 분명하다. 그러나 그에 못지않게 중요한 것이 개발과 사용에 따른 문제, 특히 피해 발생 시에 그 책임소재를 따지고, 피해회복을 위한 규율체계를 갖는 것이 중요하다. 인공지능의 자율적 능력이 고도화됨에 따라 인공지능에게 독자적인 법인격 혹은 전자인간(electronic person)으로 인정할 수 있을지를 따져보아야 한다는 주장이 없지 않으나, 이것은 현실과 먼 이야기일 뿐이다. 인공지능은 인간이면 쉽게 이해하는 중력의 법칙을 이해하지 못한다. 테이블에서 숟가락에 왜 땅바닥으로 떨어지는지를 모른다. 최고 수준의 인공지능 알고리즘이라 해도 자연어를 이해하지 못한다. 그저 흉내 낼 뿐이다. 단어 간의 연결관계를 매핑해서 의미를 추출할 뿐인데, 이를 두고 인공지능이 자연어를 이해하는 것으로 착각해서는 아니 된다. 인공지능에게는 우리 인간 통찰의 근원인 이해력은 아예 없다(Hofheinz, 2018).

2. 규율의 차등화

규제가 필요하다고 하더라도 그 기본 원칙에서는 동일하겠지만, 실제 적용과정에서는 그 규율 영역에 따라 규제의 정도나 방법이 달라져야 한다. 인공지능 적용영역은 크게 공적 영역과 사적 영역으로 나눌 수 있고, 공적 영역의 특징은 누구에게나 적용이 되는 반면, 사적 영역의 경우 일부 예외를 제외하고는 적용 대상이 한정되고, 선택적일 수 있다. 암 진단 기능의 닥터 왓슨의 도입은 2016년 지방대 병원을 중심으로 반짝 특수

를 누렸다가, 그 이후 추가 도입이 되지 않았다. 암 진단에 있어 닥터 왓슨의 정확도가 떨어진다는 점이 작용했겠지만, 서울의 대형병원에서는 환자들이 기존 의료진을 신뢰하고, 굳이 생소한 인공지능에 의한 진단을 선택하지 않았기 때문이기도 하다. 이와 같은 의료 진단 알고리즘의 경우는 인공지능 알고리즘을 선택하지 않고, 인간 의료진의 판단을 선택할 수 있는 선택지가 있기 때문에 그러한 분야의 규율은 공적 영역에서의 규율과는 차등을 둘 수 있다. 공적 영역의 경우, 예를 들면 피고인의 재범 위험성 예측 알고리즘은 적용 대상자들이 그 알고리즘 대신 선택할 수 있는 선택지가 없다. 무조건 그 알고리즘의 결과에 따라 석방 여부 등이 결정된다. 또 인공지능 알고리즘이 물리적인 영역과 결합되지 않으면 인간의 생명이나 신체 등에 직접적인 위험을 야기하지 않는다. 그런 경우 주로 차별이나 편향성 등이 문제되지만, 자율주행자동차와 같이 알고리즘이 물리적 영역과 결합된 경우는 그 알고리즘이 문제가 있으면 사람의 생명이나 신체 등이 직접적인 위해의 결과를 초래한다. 따라서 그 규율의 강도나 완전성에서 차이를 두어야 한다. 특히 인공지능 알고리즘이 위협적인 무기 등과 결합될 때에는 위협적인 무기 자체를 규율하는 기존 법률보다 강화된 어떤 규제가 필요하다. 기존의 위협적인 무기 시스템이 인공지능이라는 신기술과 결합되는 순간 기존의 무기와는 다른 이질적이고 더 위험한 존재가 되어 기존의 무기에 대한 법규로는 규제가 어렵기 때문이다. 신약 개발 단계에서 여러 단계의 임상실험을 거치는 까다로운 허가 과정을 거쳐야 하는 것은 그런 신약이 잘못되었을 때 그 부작용이 사람의 생명이나 건강에 직결되기 때문이다. 이런 규율의 차등화를 고려하지 않고, 모든 영역에 통하는 규제 원리라는 것이 실효성이 없을 수밖에 없다.

여기에 부가할 것은 기기와 결합되지 않는 알고리즘 자체도 위험할 수 있다는 점이다. 특히 차별성이나 편향성을 조장할 수 있는 알고리즘은 그 사용결과가 인류에게 소중한 여러 가지 가치를 훼손하기 때문이다. 인간의 행태에 대한 분석을 통해 인간의 성향 등을 분석해내는 알고리즘은 인간의 존엄성 자체를 훼손할 수 있는 후속적 조치로 연결되는 위험성을 가지고 있다.

3. 사전 예방적 규율

자율적인 윤리 대응도 사전 예방책의 일환일 수 있지만, 엄밀한 의미의 사전 예방책은 강제력이 있는 법적 규제책이어야 한다.

인공지능이 가지고 있는 위험성을 해소하기 위해서는 안전성 규제가 최우선 과제다. 개발 단계에서도 안전성을 확보할 수 있도록 개발과정에 대한 보고의무 부과나 안전성 심사 등의 규제방안이 필요하며, 개발 착수 전에 개발 허가과정의 심사를 통하여 개발하고자 하는 인공지능 시스템의 설계 내용이 각종 관계법령이나 안전기준을 충족하는지를 평가하는 방안도 생각할 수 있다.

또 사용 승인 후 일정 기간 사용과정을 사후 평가하여 안전 중요도가 높은데도 그 기준 충족이 미흡하다고 판단될 경우, 생산자나 최종사용자에게 인공지능 시스템의 실질적 안전성을 향상시키는 것과 같은 장기적 안전성 유지대책 방안을 강구토록 하는 것도 필요할 것으로 보인다. 또 최종적 사용승인 전에 신약허가와 같이 일정한 형태의 사용 실험을 통해 안전성 여부를 충분히 검증한 후, 그런 검증을 통과하여야 비로소 사용승인을 하는 시스템을 통해서도 사전 규제가 가능하다. 또 무인자동차의 경우 그 사용에 있어서 사용자에게 별도의 자격증 획득을 요구하거나, 그 자동차를 주간에만 사용할 수 있도록 제한하기도 한다. 보험 요구도 일반 자동차와는 다르게 하여야 한다.

우리나라의 경우에도 인공지능 시스템이 가지는 위험성에 비

추어 이와 같은 안전기준을 정하고, 준수하도록 하는 입법의 필요성은 당연하다. 또 인공지능 시스템의 운영 경험이나 그러한 시스템이 초래된 위험 유형에 대한 체계적 관리를 통해, 향후 인공지능 개발 규제의 지침으로 활용할 수도 있다. 나아가 국제기구나 협의체를 중심으로 인공지능의 안전기준 공조를 위한 노력도 필요하다.

4. 법적 책임 귀속을 통한 사후적 규율

인공지능의 법적 책임과 관련하여 현실적인 해결책은 일단 인공지능의 특질을 고려하는 것이다. 인공지능이 가진 특질을 고려하지 않을 경우 그 사용으로 인한 결과에 대하여 일반 제조물과 같이 제조자, 판매자, 소유자 중 한 사람에게 책임을 지우는 것은 책임법리상 당연한 것으로 보인다. 책임의 주체에 관한 논의에서 제조자를 책임주체로 떠올리는 것은 자연스럽다(Marchant & Lindor, 2012). 제조자가 손해를 야기한 디자인의 결함을 알았거나 알 수 있었을 때 제조자는 그 손해에 대하여 책임을 져야 한다(Belay, 2015).

인공지능 개발자 또는 제조자가 아니라면 인공지능의 소유자 또는 사용자가 책임을 져야하는 것이 순서상 맞다. 그러나 통상 이러한 소유자나 사용자가 인공지능의 의사결정과정에 아무런 역할을 하지 않기 때문에 기존의 과실 책임 논리로는 이들의 책임에 대한 설명을 흡족하게 하지 못하다. 이런 책임 불명의 문제에 지나치게 천착하면 인공지능 윤리 규범 문제를 개발이나 제조단계가 아닌 사용 단계에서 조정한다는 발상까지 하게 된다. 자율주행자동차의 사용자는 자율주행자동차의 윤리적 우선순위를 조정할 수 있도록 해야 한다는 논리가 바로 그것이다(Belay, 2015). 그러나 인공지능의 작동 과정은 사용자조차 통제할 수 없는 상황이 있으므로, 그러한 단순한 책임논리를 따를 수 없다. 그에 대한 대안을 고려할 때, 수많은 컴포넌트로 구성된 인공지능 시스템의 특성 때문에 누가 개발한, 어떤 컴포넌트 때문에 결함이 생겼는지 파악하는 것이 쉽지 않다는 점이 우선 고려되어야 한다. 설사 인공지능 알고리즘 개발과정에서의 설계상의 결함으로 인해 사고가 발생했다고 하여도, 인공지능 알고리즘의 복잡성과 불투명성 때문에 그 책임 규명과정에서 피해자는 물론 이러한 인공지능 시스템의 최종사용자, 컴포넌트 개발에 관여한 자, 인공지능 시스템 판매자나 시스템 관리자 등 어느 누구도 그 결함여부를 쉽게 발견할 수 없다. 설계나 변경, 컴포넌트의 개발, 조립 등 인공지능 개발에 다양한 많은 사람이 여러 단계에 걸쳐 관여하였기 때문에 이러한 주체들의 책임범위를 구체적 현실상황에 맞게 밝혀내는 문제는 만만한 문제가 아니다. 이러한 불투명성과 복잡성 때문에 인공지능 시스템 사용으로 인해 제3자에게 야기한 손해에 대한 책임을 소유자 혹은 최종사용자에게 묻는 것도 쉽지 않을 것이다(Scherer, 2018).

인공지능 시스템의 경우, 설계자의 의도대로 제조되어도 안전하지 못하고, 손해가 발생하는 경우가 있기 때문에 현행의 제조물책임법에 의하여 규율할 수 없다. 인공지능으로 인한 손해 발생의 주원인이 어디에 있는지, 손해의 발생이 예견가능한가 하는 일반 기준은 인공지능 시스템에 적용하기 어렵다. 따라서 인공지능의 특성을 반영할 별개의 책임요건을 규정할 필요가 있다.

끝내 특정인에게 책임을 물을 수 없는 상황이라면 인공지능으로 인한 피해자 모두에게 혜택이 돌아가는 개보험(皆保險) 제도의 도입도 고려해볼 필요가 있다.

현행법의 체계 내에서 굳이 책임 문제를 해결하고자 한다면 앞서 논의한 인공지능 윤리 지침과 관련하여 다음과 같은 방안을 강구하는 것도 하나의 선택지다. 즉 인공지능 시스템의 경우, 그 개발자나 설계자가 윤리 지침을 제대로 준수하지 않고, 생명 · 신체에 위해를 발생시킬 위험이 있는 인공지능 시스템을 설계하여 제조 · 판매한 경우에는 특별한 사정이 없는

한 그 인공지능 시스템에는 사회통념상 통상적으로 기대되는 안전성이 결여된 설계상의 결함이 존재하고, 따라서 그 결함에 대한 책임을 개발자가 져야 한다는 방안이다. 나아가 머신러닝 알고리즘의 불투명성, 예측 불가능성 등의 여러 가지 속성은 널리 알려진 위험 요소이고, 그런 위험을 잘 알면서도 그런 알고리즘을 개발을 감행하고 실제 운용하기에 이르렀다면, 인공지능이 야기한 손해에 대한 책임을 개발자나 사용자에게 물어야 한다.

이러한 민사적 책임과 달리 형사적 책임의 귀속은 인공지능의 주체성 논의와 맞물려 복잡한 양상을 띤다. 예를 들어 마이크로소프트의 챗봇(chatbot)이 채팅과정에서 사람을 모욕한 경우를 가정해보자. 그 챗봇에게 모욕죄의 죄책을 묻기는 어렵다. 물론 사람에게 책임을 지운다고 할 때, 어떤 사람이 의도적으로 이러한 챗봇을 이용하여 타인을 모욕한 경우임이 밝혀지면, 챗봇을 도구로 간주하고, 그것을 이용한 사람에게 형사책임을 지우면 된다. 그렇지 않은 경우라면 인공지능 알고리즘의 제조사 또는 소유자의 주의의무위반을 이유로 책임을 묻는 방안이 있을 수 있다. 이런 경우도 주의의무 위반을 처벌하는 법규가 있고, 제조사나 소유자에게 설계를 지침을 따르지 않은 윤리 위반이 있는 경우, 이를 주의의무 위반의 과실이 있는 것으로 보아 처벌이 가능할 수도 있다. 물론 위 사례는 과실에 대한 벌칙이 없어서 처벌하지 못한다. 그러나 인공지능이 자발적인 행동으로 손해를 발생시켰고, 인간이 전혀 예상할 수 없었던 경우에는 직접적 관여가 없었던 인간에게 책임을 지울 수는 없다. 그러나 이와 같은 문제점이 있다고 하여 인공지능에게 ☒전자인간(electronic person)' 지위를 부여하여 형사책임을 묻는 방안을 모색하거나, 심지어 제작 시 장착한 ☒킬 스위치(kill switch)'를 사용하여 인공지능을 정지시키는 방안을 모색하는 것은 무의미하다. 킬 스위치는 다시 가동하면 되는 것이고 나아가 인공지능 로봇을 파괴한다고 한들 그것을 인간에 대한 형벌과 동일시할 수 있는 것도 아니다.

참고문헌

Thomas Burri. (2016). Machine Learning and the Law: Five Theses. NIPS Symposium, 8 482–494.

https://doi.org/10.1016/j.cnsns.2016.08.011.

Corinne Cath. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans A Math Phys Eng Sci. 28;376(2133).

J.P. Gunderson&L.F. Gunderson. (2016). Intelligence ≠ Autonomy≠ Capability.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.8279&rep=rep1&type=pdf

Bostrom, Nick & Yudkowsky, Eliezer. (2011). The Ethics of Artificial Intelligence, Cambridge Hanbook of Artificial Intelligence, William Ramsey and Keith Frankish(eds.), Cambridge University Press.

Leikas, Jaana & Koivisto, Raija & Gotcheva, Nadezhda. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. Journal of Open Innovation: Technology, Market, and Complexity. 5. 18. 10.3390/joitmc5010018.

Crawford, K. (2016). Artificial Intelligence's White Guy Problem. The New York Times.

https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.

Kirchner, J.& Angwin, S. & Mattu, J. &Larson, L. (2016). Machine Bias: There's Software Used across the Country to Predict Future Criminals, and It's Biased against Blacks. Pro Publica: New York, NY, USA.

Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety. The Alan Turing Institute

Morris, D. Z. (2016). U.N. Moves Towards Possible Ban on Autonomous Weapons.

https://fortune.com/2016/12/24/un-ban-autonomous-weapons/

Salon Barocas & Andrew D. Selbst(2016). Big Data's Disparate Impact. 104 Cal. L.Rev. 671 2016.

Paul Hofheinz. (2018). The Ethics of Artificial Intelligence. The Lisbon council, interactive policy brief issune .

Pavaloiu, A., & Kose, U. (2017). Ethical Artificial Intelligence - An Open Question. Journal of Multidisciplinary Developments. 2(2), 15-27

Gary E. Marchant & Rachel A. Lindor. (2012). The Coming Collision Between Autonomous Vehicles and the Liability System. 52 SANTA CLARA L. REV. 1321, 1329.

Nick Belay. (2015). Robot Ethics And Self-driving Cars: How EThical Determinations In Software Will Require A New Legal Framework. 40 J. Legal Prof. 119.

Matthew U. Scherer. (2016). Regulating Artificial Intelligence System: Risks Challenges, Competencies, And Strategies. Harvard Journal of Law & Technology Volume 29, Number2.

Matthew Sparkes. (2006). The immoral machine. IET Computing & Control Engineering.