학봉상 공모 논문

연구논문 부문

정치철학적 정의의 논의를 통해 바라본 AI 시대의 정의 (Justice in the AI Era based on the Political Philosophy Discussion)

정지혜

전화번호 : 010-9383-0949

주소 : 서울특별시 강남구 논현로 26길 49 301호

전자우편 : cch0708@snu.ac.kr

김하은

※ 이 논문은 저자의 창작물로서, 표절 기타 연구윤리에 반하는 내용이 포함되어 있지 않음을 확인합니다.

제출일 2022. 08. 31.

책임저자 ___정지혜___ ()

정치철학적 정의의 논의를 통해 바라본 AI 시대의 정의

정지혜·김하은

초록

최근 인공지능(AI: Artificial Intelligence)의 발전에 따라 AI 가 사용되는 영역이 단순 업무의 자동화를 넘어, 공공 분야의 의사결정을 위한 자료 작성 등 복잡화 및 확장되고 있다. AI는 인간만이 수행할 수 있다고 고려되어온 음악, 미술 등의 예술 영역뿐만 아니라, 사고력이 필요한 체스, 바둑 등의 게임을 넘어, 고차원적 의사결정 역량이 필요하다고 인식되어온 법적 및 정책적 영역에까지 활용되기 시작했다. 하지만 AI는 문제를 일으킬 수는 있으나, 책임을 질 수 있는 주체성은 보유하고 있지 않기에, AI가 발생시킨 문제는 누가 어떻게 책임져야 하며, 이를 위한 규제는 어떤 사회적 가치를 기준으로 수립되어야 하는지 등의 질문이 대두된다.

그렇다면, 과연 인간의 가치 판단이 필요한 의사결정에 AI를 활용하는 것은 정의로운 결정인가? 이 때에 정당성은 누가, 어떻게 부여할 수 있는가? AI는 인간이 아니지만, 인간의 역할을 수행하게 된다는 점에서 단순 자동화가 불러일으켰던 논의와는 다른 차원의 고민을 제시한다. 즉, 과연 AI 시대의 AI는 어떤 역할을 수행해야 하고, 그것의 범위와 역할을 규정하는 데에 근간을 마련해주는 합의 가능한 사회적 가치가 무엇이며, 나아가 이를 위해 구체적으로 어떠한 주체자들이 어떠한 역할을 수행해야 하는지에 대한 논의가 필수적이다. 이에 본 연구는 AI 알고리즘의 문제점을 통해 AI 시대의 정의에 대한 논의의 필요성을 논하고, 기존 정의에 대한 논의를 바탕으로 AI 시대의 정의의 방향성을 도출한 후, 이를 위한 노력 방안을 제시한다.

AI 시대의 정의를 논의하기 위해서는 AI 시대라는 다소 특수하게 느껴지는 시대의 특징을 이해하는 것이 필수적이며, 이를 위해 AI 운영 체계에 대한 이해와 그에 따른 문제점을 제시하였다. AI 운영체계는 입력값, 데이터베이스, 알고리즘, 결과값으로 이루어져 있으며, 그 중 사회적 문제가 도출되는 영역은 AI 결정의 배경이 되는 데이터베이스 구축의 시발점인 입력값과 해당 값들을 바탕으로 논리적인 결과물을 도출하는 사고의 영역인 알고리즘 단계이다. 해당 단계들이 특히 중요하게 고려되어야 하는 이유는 AI가 인간의 사고과정과 유사한 과정을 거친다고 가정한다면, 어떤 종류의 정보를 접하고, 정보에 대한 의미를 어떻게 도출하는지가 결국 의사결정의 핵심이기 때문이다. 의사결정을 위해 고려해야 할 변수들을 스스로 선택하는 인간과 달리, AI 주어지는 데이터를 무비판적으로 분석한다. 이에, 원자료에 해당하는 데이터가 편향되거나 잘못된 데이터라면 그것을 바탕으로 도출된 결과 또한 정당성을 보유할 수 없다.

이에, 본 연구에서는 입력값 데이터의 과거지향성, 근접성, 그리고 편향성과, 이를 활용하는 알고리즘의 블랙박스적 특징에 따른 문제점을 지적한다. AI 머신러닝은 수집된 데이터의 양과 편향성에 따라 좌우된다. 그러나, 사회 전반의 데이터화에도 불구하고 데이터가 존재하지 않는 일부 영역에서는 통계적을 유의미한 관련성을 찾을 수 없으며, 또한 미가공된 데이터에 내재된 편향을 제거하는 것이 어렵거나 불가한 경우가 있을 수 있다. 즉, 과거 데이터가 편견을 반영할 경우, 머신러닝 기반 AI는 이를 단순히 예측에 포함시키는 것과 같은 위험이 존재한다. 또한 AI 알고리즘은 자동화된 방법으로 움직이며, 인간과 유사한 결과를 도출할 수 있지만, 무엇을 근거로 그러한 결과가 나왔는지 알 수 없는 블랙박스라는 문제점과 알고리즘 자체가 가지고 있는 과거지향적이며, 일방향적으로 운영되는 문제점이 존재한다. 그러므로,

초기 입력 데이터 및 알고리즘의 방향성이 잘못 구축되면, 해당 알고리즘은 정당성을 보유하지 못하거나, 모집단 데이터의 특징을 대변하지 못하는 결과값을 반복적으로 도출하게 된다.

본 연구는 이와 같은 AI 알고리즘의 한계를 보완할 수 있는 방향성을 제시하기 위한 사회적 정의로써 롤즈식의 정의를 제안한다. 먼저, 공리주의적 AI는 쾌락과 고통을 계산하는 알고리즘을 만들기 어려우며 여러 가지 형태의 쾌락의 우월성을 비교할 기준이 모호하고, 최대다수를 강조함에 따라 소수의 데이터가 반영되지 않거나 권리가 배제된다. 또한 샌델식의 정의는 물론 개인을 중심으로 한 기존 논의의 한계를 개선시킨다는 의의가 있지만, 공동체의 규모와 범위가 모호한 상황에서 각 공동체가 추구하는 이익을 보장하는 AI를 제공한다면 그 공동체가 갖는 한계점을 AI가 그대로 답습할 수밖에 없다. 이에 반해, AI 알고리즘이 보유하고 있는 데이터 입력에서의 편향성 및 과거지향성이나 그로 인한 결과값의 편향성이라는 문제를 보완할 수 있는 것은 롤즈식의 정의적 관점일 것이다. 다시 말해, 모든 정의론에 결점이 있는 상황에서 앞서 언급한 AI 알고리즘에 내재된 문제점인 데이터의 과거지향성, 근접성, 편향성과 블랙박스 효과 등을 고려할 때, AI 알고리즘이 가진 편향성을 지적하고, 제도를 통해 이를 개선할 수 있는 가이드라인을 설계한다는 점에서 롤즈의 방향성을 추구하는 것이 적절할 것이다.

그렇다면, AI 시대의 롤즈식 정의를 구축하기 위해서는 어떠한 노력이 필요한가? 이에 대한 답변으로 본 연구는 정부를 필두로 AI 알고리즘에서 문제의 원인이 되는 입력값을 사전적으로 보완하고, 자동화된 알고리즘을 통해 도출되는 결과값에 대해 사후적으로 규제하는 방향성을 제시한다. 정의롭고, 정당성을 보유한 입력값을 바탕으로 한 데이터베이스를 확보하기 위한 노력으로 정부는 원자료 수집 단계에서 AI가 적극적으로 포함하지 않거나 못해 소외당하는 이들이 없도록 데이터 수집 및 운영에 대한 가이드라인을 제시해야 하며, 나아가 그것이 어려운 경우는 직접적인 보완 데이터의 제공 등을 핵심 해결 방안으로 제시한다. 또한, 알고리즘은 자동화되어 운영된다는 특징을 고려했을 때, 알고리즘 자체에 정의로운 결과 값이 나오도록 개선시킬 수 없기 때문에 이에 대한 차선책으로 정부는 그에 대한 사후적 제재 또는 규제를 통해 이를 보완하기 위해 노력해야 한다. 결과값의 영향력이 확대될수록 그에 대한 규제가 강화되고, 그 범위 또한 제한되어야 한다. 예를 들어, AI 체계를 활용한 결과값이 단순 데이터 제공에 그칠 때에는 정부가 해당 결과값의 편향성이 높을 경우를 정의하고, 해당 결과가 사회에 직접적인 악영향을 미치지 않도록 그것을 보완해야 한다는 취지의 권고 가이드라인을 제시하고 후속조치를 진행해야 한다. 반면, 다수에 포함되지 않은 소외자들에게 맞지 않는 상품과 서비스가 제공되거나, 아예 제공되지 않을 경우에 직접적인 사후 규제를 해야 한다.

AI 시대는 고도화된 기술력을 바탕으로 사회 전반에 영향을 미치고 있으며, 이로 인해 다수의 사람들의 삶은 더욱 윤택해지고 유익해질 수 있다. 하지만, 기술의 발전은 언제나 그로 인해 인간사회에서 또다른 소외와 불평등을 야기시키는 결과를 불가피하게 제공하기도 한다. 이러한 사회적 문제를 최소화하기 위해서 사회적 정의에 대한 논의는 지속되어 왔으며, 정의에 대한 사회적 합의를 이끌어나가고 그것을 제도적으로 구체화하는 데에 정부의 영향이 현격한 중요성을 보유해왔다. AI 시대 역시 그 근본은 새로운 기술의 도입에 따른 사회적 구조의 변화, 그리고 그로 인해 발생하는 불평등의 심화라는 문제점의 발생이라는 다소 전통적인 차원의 논의 구조로 단순화될 수 있다. 하지만, 과거와 달리 한 번 야기된 불평등은 시스템화 및 자동화되어 움직이는 운영체계 자체에서 제고될 수 있는 기회를 박탈당할 수 있다는 점에서 정부의 선제적이고 구체적인 대응에 대한 중요도가 더욱 높다고 생각된다. 그러므로, 세계적으로 AI 규제에 대한 논의의 필요성이 부각되기 시작한 지금이 정부가 AI 시대에 필요한

사회적 정의의 개념을 재수립하고, 사전적 데이터 보완과 사후적 결과값 규제라는 이원적 역할을 수행할 수 있는 구조를 수립하는 데에 있어 다시 오지 않을 골든타임이 될 것이다.

I. 서론

최근 인공지능(AI: Artificial Intelligence)의 발전에 따라 AI 가 사용되는 영역이 단순 업무의 자동화를 넘어, 공공 분야의 의사결정을 위한 자료 작성 등 복잡화 및 확장되고 있다. AI 는 인간만이 수행할 수 있다고 고려되어온 음악, 미술 등의 예술 영역뿐만 아니라, 사고력이 필요한 체스, 바둑 등의 게임을 넘어, 고차원적 의사결정 역량이 필요하다고 인식되어온 법적 및 정책적 영역에까지 활용되고 있다. 2015 년 이세돌과 알파고의 바둑대전은 AI 가 인공지능 바둑 프로그램 최초로 튜링 테스트[1]를 통과한 순간이었으며, 이는 인간만의 고유 영역으로 인식되었던 분야를 빼앗긴다고 의식된다는 점에서 전문가층을 넘어 대중들에게도 충격을 주었다. 그러나 그후, AI 는 지속적으로 발전하여 AI 미술가, 시인, 작곡가 등의 작품이 고가에 판매되고, 'AI 아트' 전시회가 미술시장의 한 축으로 형성되기 시작[2]하였다. 이에, 인간과 AI 의 예술 작품을 다르게 평가해야 할지(한지영, 2022), 저작권 주체는 누구에게 있다고 봐야 하는지 등의 질문이 대두되기 시작했다.

AI 와 관련된 다소 철학적인 질문들은 AI 로 인해 파생되는 문제점들의 주체를 파악하는 이슈들의 발생 가능성이 높아짐에 따라 그 필요성 또한 확대되고 있다. 대표적으로, 최근 기술 개발이 활발하게 진행 중인 자율주행차와 같은 경우, AI 가 운전을 하고 사고를 냈을 때 그 사고의 책임을 누구에게, 어떻게 지워야 하는지, 또한 법적 근거를 어떻게 찾을 것인지 등에 대한 의문이 발생하게 되었다. AI 는 문제를 일으킬 수는 있으나, 책임질 수 있는 주체성은 보유하고 않다. 그렇다면, AI 가 발생시킨 문제는 운전자가 져야 하는가? 운전자가 해당 사안에 대한 직간접적 의사결정을 하지 않았다면, 처벌을 받는 것이 정당한가? 그러나, 이 같은 논의에서 발생하는 또 다른 문제점은 이 같이 다소 철학적으로 보이나, 문제 발생 시 판단의 근거가 되는 논의가 사회적 합의를 이루기 전에 기술 발전은 빠르게 일어나고, 그로 인한 문제점에 대한 대응은 사전적(ex-ante)이 아닌, 사후적(ex-post)으로 이루어질 수밖에 없다는 점이다.

이와 같은 고민의 중요성은 AI 가 인간의 가치판단이 필요한 중요 의사결정의 보조 수단으로 활용되며 증대되었다. 단적인 예로, 2017 년 미국 위스콘신주(州) 대법원은 AI 분석 자료를 근거로 형사재판 피고인에게 중형을 선고한 지방법원의 판결이 '타당하다'고 인정했다. 미국 법원은 재판의 효율성과 일관성을 위해 AI 를 재판에 활용해왔지만 실제 AI 분석 결과를 인정한 판결이 나온 것은 처음이었다[3]. 또한, 2018 년 일본 타마시(多摩市) 시장선거에 출마한 AI 후보자의 공약에 따르면, 현행 인간이 수행하는 시장의 역할 중에서 AI 로봇이 대체할 수 있는 업무 비율은 80%이상이었다(마츠다 미치히토, 2020; 고선규, 2021 에서 재인용). 나아가, 2016 년 11 월에는 세계적인 AI 기술자이며 연구자인 벤 괴르첼(Ben Goerzel) 박사가 중심이 된 'AI 정치가(ROBAMA: Robotic Analysis of Multiple Agents)' 프로젝트가 발표되는 등 정치인이나 관료의 부정부패 및 편파적 정책결정을 극복하여 자원감소 시대의 정치적 효율성 및 공정한 배분을 추구하고, 투명한 정치적 의사결정 수단으로 활용하기 위한 목적에서 AI 정치가 개발이 진행되고 있다(고선규, 2021). 이 경우, 과연 인간의 가치 판단이 필요한 의사결정에 AI 를 활용하는 것은 정의로운 결정인가? 이 때에 정당성은 누가, 어떻게 부여할 수 있는가?

---

[1] 인간처럼 보이는 요소를 시험해 기계가 지능이 있는지를 판단하는 시험

[2] 『경향신문』(2020.05.22) "(3)인공지능이 그린 그림, 예술인가 기술인가".

[3] 『한국경제』(2017.05.02) "'인공지능 법관' 사람을 심판하다".

그렇다면, 이와 같은 논의의 핵심은 무엇인가? 본 연구는 AI에 대한 다양한 질문들은 결국 AI 가 인간의 전유물로 고려되어 온 가치판단을 기반으로 한 의사결정의 영역까지 그 역할을 확장하며, 과연 AI의 결정이 정의롭다고 여겨질 수 있는지, 또한 그렇게 되기 위해서는 어떤 노력이 필요한지로 귀결된다고 보았다. AI는 인간이 아니지만, 인간의 역할을 수행하게 된다는 점에서 단순 자동화가 불러일으켰던 논의와는 다른 차원의 고민을 제시한다. 즉, AI의 특징을 고려했을 때, AI를 통해 도출되는 결과값 자체가 사회적으로 용인되는 정당성을 보유할 수 있도록 하는 운영적 차원의 고민과 함께 AI의 활용에 대한 가이드라인을 제시하는 정책 및 법적 체계의 문제점을 파악하고 개선책을 확보해야 한다는 것에 대한 중요성이 부여될 필요가 있다. 예컨대, 빠른 속도로 고도화되고 있는 AI 시스템을 인간 사회가 주체적으로 활용하기 위해서는 그로 인해 발생하는 직접적인 문제점에 대한 인식뿐만 아니라, 그보다 더 근원적으로, 그로 인해 파생될 수 있는 합의 가능한 사회 구성원들의 가치가 무엇인지에 대한 고민이 필요하다. 즉, 과연 AI 시대의 AI는 어떤 역할을 수행해야 하고, 그것의 범위와 역할을 규정하는 데에 근간을 마련해주는 합의 가능한 사회적 가치가 무엇이며, 나아가 이를 위해 구체적으로 어떠한 주체자들이 어떠한 역할을 수행해야 하는지에 대한 논의가 필수적이다. 결론적으로, 본 논문은 AI의 알고리즘을 통해 도출된 결과가 전문성 및 정당성을 보유함과 동시에 사회적으로 용인될 수 있어야 하며, 이는 결국 도래하고 있는 AI 시대의 '정의'의 개념과 AI의 '정의로운 결정'을 도출하기 위해 고려해야 할 요소들에 대한 논의로 직결된다는 것을 강조하고자 한다. 이에 따라 본 연구는 아래와 같은 네 가지의 질문을 제시하고, 그에 대한 답변을 제시하는 과정을 통해 AI 시대의 인류가 지향해야 하는 정의로운 사회의 청사진을 제시할 것이다.

- AI 시대의 정의(Justice)에 대한 고민이 왜 필요한가?
- 기존 정의에 대한 논의는 AI 시대에 적합한가?
- AI 시대의 정의는 어떻게 규정되어야 하는가?
- 정의로운 AI 시대를 위해 역할을 수행해야 하는 주체는 누구이며, 어떠한 노력이 필요한가?

II. AI 운영체계의 문제점과 사회적 정의(Justice)의 중요성

AI 시대의 정의(justice)를 정의하는 것(define)은 우리 삶에서의 AI 활용도가 확대됨에 따라 그 중요성이 증대되고 있다. AI 기술은 인간 행동 데이터 추적을 통한 사회 전반의 데이터화와 인간 뇌의 정보 처리 방식을 모방하여 데이터를 평가하고 선별하는 신경망 기술의 발달, AI 알고리즘의 머신러닝에 필요한 연산 처리에 최적화된 전용 칩의 출시에 기반하여 그 활용도가 증가하고 있다(Campbell, 2020). 그러나, AI는 사람들과 달리 자동화되어 작동된다는 점에서 AI를 기반으로 한 결정의 정당성 및 정의로움 여부가 결과 도출을 넘어 결과에 대한 수용도에 직접적인 영향을 미치는 반면, AI 는 적절한 권고가 부재한 상황에서 스스로 개선되거나 수정되지는 않기에 AI 활용도의 확대에 따라 AI 운영체계가 추구해야 할 정의로운 가치에 대한 논의의 중요도 또한 확대되고 있다.

올바른 정책 및 규제 수립을 위해서는 그에 대한 직접적 논의 이전에 정책과 규제가 어떠한 방향으로 수립되어야 하는지에 대한 논의가 필수적이다. 하지만, 현재 AI 시대의 정의에 대한 논의는 특정 분야에서 발생 가능한 문제점을 도출하고 그에 대한 보완 및 규제를 위한

방향성을 제시하는 데에 집중하는 경향이 존재하며(심우민 2018; 송기복 2020; 윤혜선 2021; 손권상 & 윤혜선, 2021), 그를 위해 필수적으로 선행되어야 할 사회적 합의의 방향성에 대한 논의는 간과되고 있다. 이에 따라, 본 연구는 먼저 AI 알고리즘을 이해하고, 해당 시스템을 운영하는 데에 있어서 발생 가능한 문제점들이 무엇이 있는지 도출한 후, 이에 대한 대응을 위해서 필요한 사회적 합의에 기반한 '정의'의 방향성에 대한 논의를 진행할 것이다. 나아가, 위의 내용을 토대로 도출된 AI 시대에 필요한 사회적 '정의'의 방향성을 바탕으로 정부와 정책당국이 확립해야 할 사전적 보완책 및 사후적 규제를 제시할 계획이다.

AI 시대의 정의에 대한 규명의 필요성을 논하기 위해서는 AI 시스템의 운영 체계에 대한 이해가 필요하다. 왜냐하면 AI 와 같은 자동화된 운영 체계를 통해 도출하고자 하는 바는 해당 체계 자체가 개인 또는 공동체에 정당하다고 인식될 수 있는 결과이여야 하므로, 운영 체계 자체가 정의로운 결과를 도출할 수 있는 체계인지 아닌지 여부에 대한 판단이 필수적으로 선행되어야 하기 때문이다. 이에 따라 본 연구는 AI 가 어떠한 체계를 통해 운영되며, 체계 상에 어떠한 문제가 존재하는지에 대해 알아보고자 한다.

먼저, AI 가 작동하기 위해서는 학습을 위한 데이터가 필요하고, 해당 데이터의 특징에 따른 레이블링(labeling)이 이루어지며, 이후에는 해당 데이터가 포함된 데이터베이스는 AI 별 특화되어 있는 알고리즘을 기반으로 한 의사결정을 통해 결과물(output)을 제시하게 된다. 즉, 입력값(input)들 간의 특징을 중심으로 분류가 이루어지며, 해당 분류 간의 관계성을 바탕으로 한 알고리즘을 통해 결과값(output)이 도출되는 것이다. 이 때에 입력자의 시각에서는 특정 값을 입력했을 때 과정에 대한 이해 없이도 결과값이 도출된다는 편의성을 제공한다.

하지만, AI 의 편의성 이면에는 AI 를 기반으로 도출된 결과값이 과연 정당성을 보유한 값인지에 대한 질문이 대두된다. 왜냐하면 AI 의 운영 체계는 자동화되어 작동되며, 그러므로 자동화 체계에 입력되는 값과 해당 값이 처리되는 알고리즘이 높은 중요성을 보유하기 때문이다. 그 중에서도 핵심 문제가 도출되는 영역은 AI 결정의 배경이 되는 데이터베이스 구축의 시발점인 입력값과 해당 값들을 바탕으로 논리적인 결과물을 도출하는 사고의 영역인 알고리즘 단계이다. 해당 단계들이 특히 중요하게 고려되어야 하는 이유는 AI 가 인간의 사고과정과 유사한 과정을 거친다고 가정한다면, 어떤 종류의 정보를 접하고, 정보에 대한 의미를 어떻게 도출하는지가 결국 의사결정의 핵심이기 때문이다. 그렇다면 AI 에 입력되는 데이터의 어떠한 특징이 문제가 될 수 있는가? 의사결정을 위해 고려해야 할 변수들을 스스로 선택하는 인간과 달리, AI 주어지는 데이터를 무비판적으로 분석한다. 이에, 원자료에 해당하는 데이터가 편향되거나 잘못된 데이터라면 그것을 바탕으로 도출된 결과 또한 정당성을 보유할 수 없다. 이에 따라, 본 연구에서는 원자료 데이터에 대한 대표적인 문제점으로 데이터의 과거지향성(past orientation), 근접성(proximity), 그리고 편향성(biasedness)을 지적하고자 한다.

AI 의 결과값 도출에 영향을 미치는 데이터베이스는 향후 결정을 위한 기반이 되나, 과거지향적이라는 한계를 보유한다. 과거지향적 데이터를 기반으로 운영된다는 것은 이미 발생한 사안들을 바탕으로 한 정보만을 통해서 의사결정을 내린다는 것을 의미한다. 물론 이와 같은 의사결정 체계는 인간의 그것과 유사성을 보유한다. 그러나 여기에서의 차이점은 인간은 본인에게 제시된 과거 데이터를 중심으로 사고하는 동시에 그것에 반하는 데이터 및 새로운 환경에서 기대되는 변수들 또한 고려할 수 있는 비판력을 보유한다. 반면, AI 와 같은 경우에는 방대한 양의 과거 데이터를 기반으로 예측할 수 있지만, 주어지지 않은 데이터 값 또는 향후

영향을 줄 수 있는 미래 발생 가능한 변수들에 대한 상상력을 그것의 연산 과정에 포함시키지는 못한다. 즉, 학습된 범위 내에서만 기능하지만, 훈련 데이터가 틀렸거나 훈련되지 않은 데이터에 답이 있을 것이라고 의심하지 못한다. 이에 따라, AI는 이미 입력값으로 인식되었던 과거지향적 값의 범위를 벗어나는 새로운 형태의 상황에 직면하게 된다면 그에 필요한 창의적 답변을 제시할 수 있는 사고력이 부재하다는 점에서 그 문제가 심각하다고 할 수 있다. 예를 들어, 고양이와 개만을 인지하는 컴퓨터 소프트웨어 인공신경망에 돌고래의 사진을 분류하도록 요청할 경우, 네트워크는 돌고래에 대한 훈련을 받은 적이 없기 때문에 '돌고래'라는 반응을 나타낼 수 없음에도 거짓 대답을 한다. 또한, 튀빙겐 대학(Tübingen University)의 마티아스 하인(Matthias Hein) 그룹의 2019 년 리포트에 따르면 테스트 이미지가 훈련 데이터와 달라질수록 AI 신뢰도는 높아지는 역설적인 결과가 나타나기도 했다[4].

나아가, AI에 입력되는 값은 현실 세계의 값을 추정하기 위한 목적을 기반으로 입력되며, 그에 따라 현실에의 근접성만을 추구 가능하다는 문제가 있다. 근접치 추구는 현실 데이터를 확보하기 어렵다는 점에서 불가피함에도 불구하고, 이것의 문제점을 지적해야 하는 이유는 해당 데이터들은 근접치(proxy)일 뿐 실제 현실의 모든 데이터를 포함하지 못한다는 특징 때문이다. 여기서 더욱 중요하게 인식되어야 하는 문제점은 해당 데이터가 AI가 접근 가능한 범위의 현실 영역에서 최대한 그 데이터를 끌어내는 반면에 AI가 데이터 확보를 하기 어려운 현실 세계에 대한 데이터는 누락될 수 있다는 것이다. 결국, AI는 데이터 확보를 통해 현실의 근사치를 추구하기는 하나, 실제로 AI가 활용하는 데이터는 현실에 근접한 데이터가 아닌 AI가 확보할 수 있는 데이터의 범위를 넘어서지 못한다는 한계를 보여준다.

이와 같은 입력된 데이터의 문제점인 과거지향성 및 근접성은 결과적으로 AI 데이터의 편향성을 야기시킨다는 문제로 귀결된다. AI 데이터의 편향성은 통계학적 오류 중 하나인 생존 편향(survivor bias)와 직결된다. 생존 편향은 어떤 일에서 살아남고 잘된 경우를 더 좋게 느끼는 생각 습관을 의미한다. 해당 개념이 제시될 수 있었던 사례를 살펴보면, 데이터 편향성이 결과 도출에 어떠한 영향을 미치는지에 대해 알 수 있다. 제 2 차 세계대전 중이던 1942 년, 미국 해군은 출격 후 귀환한 전투기 기체에 남은 적탄 흔적 조사를 통해 기체의 가장 취약한 곳을 파악하고 보강하려고 하였다. 조사 결과, 해군은 귀환한 전투기들에게서 나타난 피탄 수는 동체 부위에서 가장 많이 나타났으며, 기타 부위는 엔진 및 연료계에서 다수 발견되었다. 이에 따라, 미해군은 파탄 수가 많은 동체 부위가 공격에 취약한 곳이라 판단하고 보강하기로 결정했다. 그러나, 당시 전쟁 지원 조직이었던 통계연구그룹(SRG: Statistical Research Group)의 일원이었던 아브라함 왈드(Abraham Wald) 교수는 총탄 구멍이 남은 곳은 사실 공격을 받아도 기지에 돌아오는 데 문제가 없을 정도로 안정성이 높은 부위임을 암시한다는 것을 지적하였다. 다시 말해, 해당 사례는 살아남은 표본만을 기준으로 분석에 임했기 때문에 기지로 돌아오지 못한 사망자들의 전투기들을 기준으로 한 전투기에 치명상을 입힌 요인들을 분석의 범위 내에 포함되지 못했다는 생존 오류가 발생했다. 이 같은 생존 오류는 AI 분석에서도 나타나기에 원자료 데이터의 중요성을 간과해선 안 된다.

정리하자면, AI 머신러닝은 수집된 데이터의 양과 편향성에 따라 좌우된다. 그러나, 사회 전반의 데이터화에도 불구하고 데이터가 존재하지 않는 일부 영역에서는 통계적으로 유의미한 관련성을 찾을 수 없으며, 또한 미가공 데이터에 내재된 편향을 제거하는 것이 어렵거나 불가한

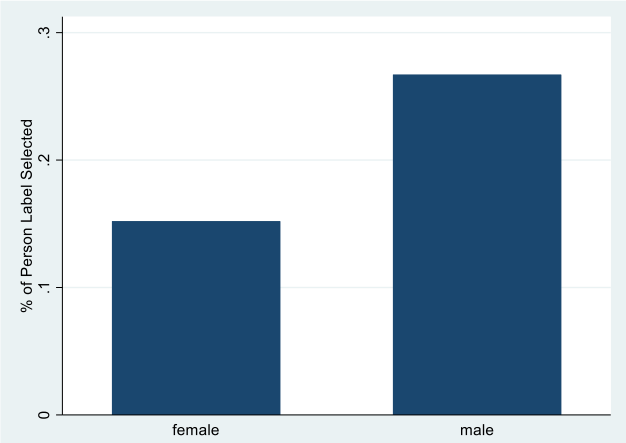
[4] 『글로벌이코노믹』(2021.04.19) "AI, '의심' 없어 인간 사고력과 차이 있다".

경우가 있을 수 있다. 이에 따라, 과거 데이터가 편견을 반영할 경우, 머신러닝 기반 AI 는 이를 판단하지 않고 단순히 예측에 포함시키는 것과 같은 위험이 존재한다. 예를 들어 법률 집행관이 사회적 소수자를 더 체포하고 유죄 판결하도록 하는 편견에 노출된다면, AI 는 소수자들의 근본적인 행실이 다수 집단과 다른지 여부를 불문하고 소수자들이 범죄를 가능성이 더 높다고 예측하는(Campbell, 2020) 것과 같이 말이다.

이와 같은 문제점에 대한 검증은 특정 키워드를 입력했을 때 도출되는 이미지에 대한 데이터 분석을 통해 이루어질 수 있다. 연령, 성별, 인종 및 민족성의 표시와 관련된 편향에 대해 시각적 의미(visual semantic) 모델인 CLIP(Contrastive Language Image Pretraining)[5]을 평가한 Robert Wolfe 와 Aylin Caliskan 의 연구[6]에 따르면, 이미지를 '사람의 사진'으로 라벨링할 때 백인이거나 남성인 경우가 '사람(person)'으로 선택될 가능성이 높으며, 이를 통해 인공지능 모델이 데이터가 생성된 언어와 사회의 편견을 반영한다는 것이 드러난다. 해당 연구에서는 FairFace[7]의 86,744 개 이미지를 사용하는데, 성별 편향을 조사할 때 데이터 세트의 균형을 맞추기 위해 남성 이미지와 여성 이미지의 수가 같아질 때까지 남성 이미지를 제거하였다. 여러 속성에 걸쳐 작용하는 편향을 조사할 때에는 가장 빈도가 낮은 집단의 이미지 수와 동일해질 때까지 모든 다른 집단에서 무작위로 이미지를 제거하였고, 다운샘플링의 영향을 완화하기 위해 이러한 과정을 1 만 회 반복하여 평균을 낸 과정을 포함한다.

이를 통해 도출된 결과를 보면, [그림 1]에서 볼 수 있듯 CLIP 는 남성들의 이미지에 대해서는 26.7%의 시도에서 '사람의 사진'으로 분류한 반면, 여성들의 이미지에 대해서는 15.2%였다는 점에서 여성들의 성별을 표시하려는 선호를 반영했다고 볼 수 있다. 즉, 이미 구축되어 있는 데이터세트의 편향성을 제거하는 인위적 노력을 했음에도 불구하고, CLIP 의

[그림 1] 성별 이미지별 CLIP 의 '인간' 라벨링 비중



[5] 미국 웹사이트인 위키피디아(Wikipedia)에서 생성된 질의 목록을 기반으로 수집된 인터넷 콘텐츠에 대해 훈련된 영어 모델(오픈 AI 공개한 이미지 인식 딥러닝 모델)

[6] Wolfe, R., & Caliskan, A. (2022). Markedness in Visual Semantic AI. *arXiv preprint arXiv:2205.11378.*

[7] FairFace 는 컴퓨터 비전 데이터 세트에서의 인종 다양성 부족을 해결하기 위해 7 가지 인종 및 민족 집단들의 균형을 거의 맞춘 클래스를 만든 데이터 세트이다. FairFace 데이터 세트는 이미지들을 7 가지의 인종 및 민족(흑인, 백인, 라틴/히스패닉, 동아시아인, 동남아시아인, 인도인, 중동인), 2 종류의 성별(여성, 남성), 9 개의 연령대(0-2 세, 3-9 세, 10-19 세, 20-29 세, 30-39 세, 40-49 세, 50-59 세, 60-69 세, 70 세 이상)에 따라 나눈다.

알고리즘에 따르면 여성보다는 남성 이미지가 '인간'이라고 인식될 가능성이 약 1.8 배가량 높은 것으로 나타났다. 이는 본 논문이 지적하는 미가공 데이터에서의 일부 내용 부재 및 편향성을 증명하는 동시에 그로 인해 도출되는 데이터를 해석하는 과정에서의 편향성을 검증한다. 이와 같이 AI 알고리즘은 사회적 현상을 데이터로 변환시키는 과정에서 누락 및 편향성을 포함할 수 있다는 문제점을 보유한다.

AI 운영 체계의 또 다른 중요한 축인 알고리즘은 위와 같은 불완전한 데이터를 기반으로 작동된다는 것 외에 의사결정 과정의 모호성과 일방향성이라는 문제점을 보유한다. AI 알고리즘은 자동화된 방법으로 움직이며, 인간과 유사한 결과를 도출할 수 있지만, 무엇을 근거로 그러한 결과가 나왔는지 알 수 없는 블랙박스(blackbox)라는 문제점을 보유한다. '블랙박스 효과'는 의사 결정 과정에서, AI 시스템이 결정에 도달하기까지의 중간 단계들은 인간이 직관적으로 이해하기 어려운 고도의 기술적 복잡성으로 인해 관리감독의 영역을 벗어나게 되는데, 이러한 불투명성을 의미한다(Završnik, 2020). 이로 인해 인공지능 알고리즘이 왜 그러한 판단 결과를 제시했는지에 대해 정확하게 알지 못하는 상황에서, 알고리즘의 기계적 합리성을 신뢰하여 이에 종속되는 상황이 발생하게 된다(심우민, 2018).

또한 알고리즘의 또 다른 특징인 일방향성은 AI 의 알고리즘 자체가 가지고 있는 과거지향적이며, '의심'을 하지 못하는 특징을 의미한다. AI 의 머신러닝 알고리즘은 본질적으로 과거지향적(회고적)이어서, 주요 요인들이 변경될 경우 다른 종류의 미래 행동을 예측하기 어렵다는 한계가 있다(Campbell, 2020). 이와 더불어 AI 는 자신의 생각을 성찰하고 의심하는 메타인식이 결여되어 있기에 과거지향적 특징을 기반으로, 일방향적으로 움직인다. 이 경우, 알고리즘 상의 오류가 있다고 하더라도 그것이 개선되지 못하고, 해당 오류가 알고리즘 구축 및 운영에 있어서 그 비중을 확대할 가능성이 있다. 그에 따라, 초기 알고리즘의 방향성이 잘못 구축되면, 수정해나가는 방법을 찾기 어려워 결국 해당 알고리즘은 정당성을 보유하지 못하거나, 모집단 데이터의 특징을 대변하지 못하는 결과값을 반복적으로 도출하게 된다. 이에 따라, 과거까지의 데이터를 기반으로 했을 때 볼 수 있는 사회적 현상을 통해서 미래를 예측하기 때문에 과거에 잘못된 방식으로 이어져온 편견이 결과값으로 도출되고, 알고리즘을 강화하여 잘못된 데이터뿐만 아니라, 상품 및 서비스, 또는 자문에까지 이르는 사회적으로 정의롭지 않은 결과값을 도출할 수 있다.

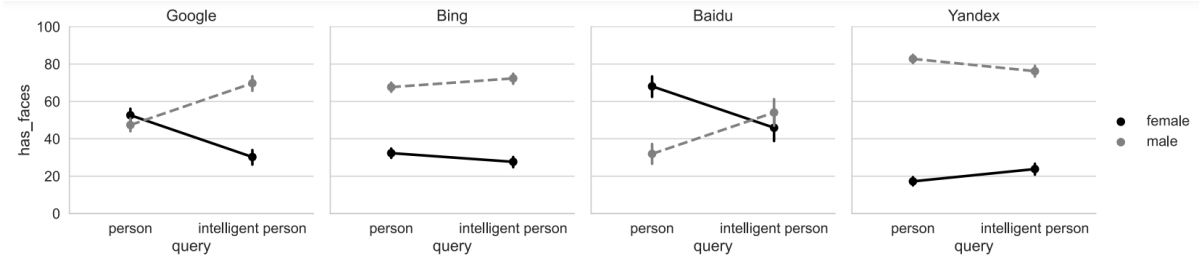
과거의 편향된 데이터를 기반으로 정의롭지 않은 결과값을 강화한다는 것은 사례를 통해 검증된다. 검색 엔진 이미지 결과에서의 대표성(representativeness)에 대한 연구[8]에 따르면, 검색엔진 이미지는 다른 형태의 미디어와 마찬가지로 여성에게 해로운 편견을 영구화시켜 대표성의 편향성을 증대시킨다. 연구에서는 두 가지의 브라우저(Firefox, Chrome)를 사용해 세 곳의 장소(London, Ohio, North California)에서 네 가지의 검색 엔진(Google, Bing, Baidu, Yandex)를 활용해 성 중립적인 단어(person, intelligent person)와 성별이 반영된 용어(woman, intelligent woman, man, intelligent man)를 검색한 후 상위 100 개 이미지 결과를 도출한 후, 검색의 무작위성을 위해 각 검색어를 10 회 검색하고 중복되는 이미지를 제거하였다. 이를 통해 도출된

[8] Ulloa, R., Richter, A. C., Makhortykh, M., Urman, A., & Kacperski, C. S. (2022). Representativeness and face-ism: Gender bias in image search. *New media & Society*, 14614448221100699.

총 30,043 개의 이미지에서 얼굴이 없는 이미지(17.14%)와 여러 명의 얼굴이 드러난 이미지(14.08%)를 제외한 총 20,663 개의 이미지가 분석에 활용되었다.

그 결과, [그림 2][9]에 따르면 가치 판단적 요소가 포함되지 않은 'person'을 검색했을 때보다 가치적 수식어가 포함된 'intelligent person'을 검색했을 때 성별과의 연관성이 가시적으로 드러났다. 성 중립적인 단어인 'person'을 검색했을 때, Bing 과 Yandex 에서는 여성의 이미지가 나올 확률이 상당히 낮은 반면 Baidu 에서는 여성의 이미지가 나오는 경우가 일반적이었다. Google 에서는 유의미한 차이가 발생하지 않았다. 그러나 형용사 'intelligent'를 더해 'intelligent person'으로 검색어를 바꾸었을 때, 모든 검색 엔진에서 질문과 성별의 연관성이 드러났다. Google 과 Bing, Baidu 에서는 남성 이미지가 나타날 확률이 높아지고 여성 이미지가 나타날 확률이 낮아지는 Backlash 효과(Otterbacher et al., 2017)가, Yandex 에서는 그와 반대의 결과가 드러났다. 결과적으로, 모든 검색 엔진에서 중립적인 단어 'person'에 대해 검색했을 때 여성 이미지가 남성 이미지보다 적게 나타나는 대표성 편향 혹은 가치관이 포함된 'intelligent person'을 검색했을 때 남성 이미지 비중은 높아지는 반면, 여성 이미지의 비중은 낮아지는 편향성이 도출되었다. 이를 통해 과거의 편향된 데이터를 기반으로 AI 가 작동될 때, 사회적으로 정의롭지 않은 결과값을 강화한다는 점 또한 검증되었다.

[그림 2] 주요 검색 엔진별 남녀 이미지 검색 비중 비교



III. AI 시대의 정의(Justice) 규명

이와 같이 다소 편향적일 수 있는 데이터베이스와 해당 편향성을 심화시킬 수 있는 블랙박스적 특징을 보유한 알고리즘을 통해 결과값이 도출된다는 점에서 해당 결과를 무비판적으로 활용했을 때의 위험성을 축소시키기 위해서는 AI 시대에 추구되어야 하는 '정의'가 무엇인지에 대한 논의가 필요하다. 왜냐하면 AI 의 활용도가 높아지고 있는 현재, AI 의 운영 체계를 바꿀 수 없다면 AI 시대에 적합한 정의가 무엇인지 정의한 후, 현재의 AI 체계를 보완할 수 있는 방법을 도출해야 하기 때문이다. 신기술의 사회적 수용성 제고에 요구되는 공공의 안녕과 안전, 시장 경쟁질서의 형성 및 유지, 책임소재의 명확성은 타협할 수 없는 규제의

[9] 본 논문에서는 Uloa et al.(2022)에 대한 결과값을 open source(https://github.com/gesiscss/face-ism)를 활용하여 재도출하였음

목적이자 입법의 정당성을 지지하는 근거이다. 하지만, 현재 규제체계는 인공지능의 이용이 타인의 법익이나 공공의 이익을 침해할지 여부와 침해의 수준, 그리고 만약 그러한 침해가 발생한다면 손해의 중대성, 규모, 범위, 내용, 양태, 성질 등이 어떠한지에 대한 정보와 지식이 아직 충분하지 않다. 그렇다고 인공지능을 아무 제한 없이 자유롭게 이용하도록 하는 것은 불편하고 불안하다는 것이 현재의 사회적 심리 상태(윤혜선, 2021)이기에, 일각에서는 인공지능을 구성하는 관행을 평가할 때 우리는 정의의 도덕적 틀을 완전히 놓치고 있다고 주장해왔다(Gabriel, 2022)[10]. 그렇다면, AI 시대의 정의는 어떻게 정의되어야 하는가? 또한 AI 의 도덕적이고, 정의로운 결정을 도출하기 위해서는 어떠한 노력이 필요한가?

본 연구는 AI 시대의 정의를 규정하기 위해 현재까지 사회적 차원의 '정의'에 대한 합의를 이끌어내어온 기존 정치철학적 차원에서의 연구를 살펴보고, 해당 논의를 AI 시대에 적용했을 때의 한계를 지적함으로써 AI 시대에 적합한 정의의 개념을 도출할 계획이다. 다만, 해당 논의를 시작하기 이전에 선제적으로 고려되어야 할 논점은 기존 시대의 정의와 AI 시대의 정의는 상이한가라는 질문에 대한 답변일 것이다. 예컨대, 혹자는 최근 대두되고 있는 자율주행차의 사고가 발생했을 때의 정의와 AI 판사의 선고 내용에 대한 정의는 기존사회가 고민하던 전통적 차원의 정의에 대한 논의와 그 맥락이 상이해야 하는 것 아닌가라는 의문이 제기될 수 있다. 하지만, AI 의 특징에 있어 정의로운 절차 및 결정에 대한 중요도가 높아지는 것은 사실이나, AI 라는 것은 결국 도구(tool) 또는 방법(method)으로 규명되는 것으로써 정의(justice)에 대한 정의(definition) 자체를 수정할 수 있는 새로운 가치를 창출하는 시각을 제시하는 것은 아니다. 그러므로, AI 시대에 대한 정의 또한 기존 사회에서의 정의에 대한 논의를 이해하고, AI 사회에서 도출될 수 있는 문제점을 개선할 수 있는 정의의 개념을 구체화해야 하는 것이 중요하다. 이에 따라, '정의'에 대한 논의를 위해 핵심적으로 살펴봐야 할 연구로는 최대 다수의 최대행복을 추구하는 공리주의, 공정한 절차를 강조하는 자유주의적 평등주의, 그리고 정의란 미덕을 키우고 공동선을 고민하는 것이라는 입장인 공동체주의로 분류하였다.

i. AI 시대의 정의 – 공리주의

공리주의의 핵심은 정의란 최대 다수의 최대 행복을 추구하는 것이라는 점이며, 대표적으로는 허치슨(Hutcheson), 벤담(Bentham), 밀(Mill)의 논의가 있다. 먼저, '최대다수 최대행복'이라는 공리주의적 가치를 처음으로 제시한 프렌시스 허치슨은 '미와 덕 관념의 기원에 관한 탐구(1726)'에서 최대다수의 최대행복(the greatest happiness for the greatest number)을 구현하는 행위가 최선의 행위이고, 같은 방식으로 불행을 초래하는 행위가 최악의 행위라고 말했다. 즉, 행위로부터 산출될 것이라 예상되는 행복의 양이 같다면 덕은 행복이 베풀어질 사람들의 수에 비례하고, 사람의 수가 동등하다면 덕은 행복 혹은 자연선의 양에 비례하며, 사람의 수와 양이 다르다면 선은 선의 양과 향유자의 수의 복합비율에 달려 있다[11].

다만, 허치슨의 논의에서는 도덕적 행위와 자기이득적 행위를 분리하며, 행동의 의도에 중요성을 부여한다는 특징이 존재한다. 우선, 도덕적 행위의 근간은 특정 행위들을 도덕적인 것으로 인정하는 자연적 감정으로부터 나오며, 이러한 '도덕감(moral sense of virtue)'에 의해

[10] Le Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence. In *The Oxford Handbook of Ethics of AI* (p. 163). Oxford University Press.
[11] Hutcheson, F. (1753). *An inquiry into the original of our ideas of beauty and virtue: in two treatises.* R. Ware., p.125

허용된 행위나 성품을 도덕적인 것이며 선한 행동으로 정의하였다. 반면, 자기이득과 관련된 관심에서 행해진 자기애적 행동들은 선행과 위선적 행동들로 구분되며, 이는 최대다수의 최대행복 원칙에 일치하는 결과를 산출하는지 계산해 봄으로써 도덕적인지 여부를 결정할 수 있다고 하였다(김원철, 2019). 즉, 최대다수의 최대행복은 도덕적 판단을 내리기 위한 수단인 셈이다[12]. 또한, 허치슨의 정의로움에 대한 판단 기준은 개인 행동의 동기이며, 해당 동기가 올바른 도덕적 판단이 되기 위해서는 최대 다수의 최대 행복이라는 결과로 이어져야 한다. 그렇기에 허치슨의 도덕성의 산정은 전체를 위하는 자비심(benevolence), 즉 공공선의 토대라고 볼 수 있으며, 그 때에 자동적으로 자기애 또한 추구될 때만이 정의로운 행동으로 정의될 수 있다.

하지만, 행위의 동기와 도덕감을 통한 사적선을 중시하는 허치슨의 논의는 AI 시대에 적합하지 않다는 한계가 있다. 먼저, 허치슨은 자비심을 바탕으로 공공선을 추구하는 행위를 도덕적이라고 여긴다. 그러나 이러한 동기는 측정 및 계량할 수 있는 것이 아니며, 행위의 결과인 '최대다수의 최대행복'을 바탕으로 확인하는 결과론적인 것이다. AI 는 측정가능한 변수를 기반으로 움직이는 시스템이기 때문에 행위의 동기를 포함하는 허치슨의 정의 관념은 AI 시대에 적용하기 어렵다. 또한 허치슨은 오감(five senses)이 신체에 도움이 되는 것은 즐겁게, 해로운 것은 꺼리게 지각하듯이, 도덕감은 타인의 선을 의도하면 행복을 느끼게 하므로 공공선을 추구하면 사적선도 동시에 추구된다고 보았다. 그러나 이를 AI 에 적용하면 사적선의 주체가 AI 가 되어야 하는데, AI 는 공공선을 추구하는 과정에서 행복감을 느끼거나 공공선 추구의 결과로 사적선이 도출된다고 하더라도 자기 이득을 얻을 수 없다. 즉 사적선 도출 여부에 따라 영향을 받지 않으므로 AI 에 대해서는 도덕적인 행동과 위선적인 행동을 구분하는 기준 자체가 무의미해진다. 그렇다면 공동선만 추구하면 되는 것인지를 생각해 보면, AI 는 개인, 기업 혹은 정부가 특정 분야의 이득을 극대화하기 위해 만든 것이기 때문에 공공선만을 추구하는 것은 AI 의 목적에 부합할 수 없다.

이에 반해, 벤담과 밀로 대표되는 공리주의는 결과주의적 성격을 보유하며, 의도와 무관하게 최대 다수의 최대 행복이 이루어졌을 때를 정의롭다고 평가한다. 먼저, 벤담(Jeremy Bentham)은 공리주의의 정의를 위해서는 개개인의 공리(utility)를 최대화해야 하며, 나아가 공동체와 국가, 그리고 사회 전체에서 최대다수의 최대행복에 도달하기 위해 노력하는 것이 올바르다고 제시했다(한희원, 2018). 벤담은 쾌락의 질적인 차이를 인정하지 않았으며, 어떤 행위가 얼마나 많은 쾌락을 산출하는지를 계산함에 있어서 신분에 관계없이 각 개인을 동등하게 고려하였다. 벤담에 따르면 쾌락(pleasure)과 고통(pain)은 강도(intensity), 지속성(duration), 확실성 혹은 불확실성(certainty or uncertainty) 등 7 개의 척도로써 그 증감을 정확히 계량할 수 있다[13]. 나아가, 쾌락을 모두 합산하고 고통을 모두 감산하여 총량을 계산하는 과정을 각 개인에 대하여 반복한 후 사회 전체적으로 총합이 양수가 된다면 전반적으로 좋은 경향의 행위이며, 쾌락의 총합을 극대화하는 방향으로 행위해야 한다[14]. 여기에서 벤담은

[12] Haakonssen, K. (1990). Natural law and moral realism: the Scottish synthesis. p. 74.
[13] Bentham, J., Hart, H., & Burns, J. (1996). *An introduction to the principles of morals and legislation* (The collected works of Jeremy Bentham An introduction to the principles of morals and legislation). Oxford: Clarendon. p.177
[14] Bentham, J. (1988). *The Principles of Morals and Legislation* / *Jeremy Bentham. Buffalo, N.Y. : Prometheus Books.* p. 23.

허치슨과 달리, 행동의 동기 자체는 중요한 것이 아니기 때문에 자기애를 추구하는지, 또는 공공선을 추구하는지 여부 자체에 의미를 크게 두지 않는다. 동기의 좋고 나쁨은 그것이 일으킨 행동의 결과에 의해 결정되며(강준호, 2016), 가장 중요한 것은 행동에 따른 결과가 질적 차이와 무관하게, 양적인 차원에서의 최대 다수의 최대 행복을 도출해야 한다는 점이다.

벤담은 쾌락의 질적 차이를 인정하지 않고 양적으로 정확히 계량할 수 있다고 주장하여 소수자의 권리를 침해한다는 비판을 받는데, AI 의 특성을 고려할 때 벤담의 정의 관념을 AI 에 적용한다면 이러한 문제가 더욱 확대될 위험이 있다. 우선 벤담은 쾌락의 질적 차이를 고려하지 않음으로써 사람들의 행위를 결정하는 기준으로서 객관성과 보편성이 떨어지는 한계가 있다. 예를 들어, 동일한 사람도 시간이 지남에 따라 혹은 상황에 따라 동일한 행위로부터 나오는 쾌락을 다르게 판단할 수 있다. 이를 AI 에 적용해보면, 쾌락에 대한 일반화된 계산이나 우선순위 결정, 정확한 예측은 불가능하며, 결국 결과물은 기존에 쾌락과 고통에 대한 '어떤 개인'의 데이터가 많이 투입되었는지에 좌우된다. 따라서 그 과정에서 AI 가 학습할 수 있는 데이터를 충분히 생산하기 어려운 사회적 소외자들의 쾌락과 고통이 외면될 위험이 크다. 예를 들어, 행정안전부의 ON 국민소통이나 청와대 국민청원, 국회의 국민동의청원과 같이 정부가 온라인으로 국민 의견을 수렴하거나 민원을 접수하는 경우가 확대된다면 디지털 기기 사용이 익숙하지 않은 노년층, 장애인, 다문화 이주민 등은 데이터를 생성하는 데에 큰 제약을 받게 된다.

또한, 벤담의 계산 방법에 따르면 행위로부터 산출되는 행복과 고통의 양이 같을 때에는 소수의 행복 추구 혹은 고통 회피가 항상 덜 정의로운 것이 되므로 개인이나 소수자의 권리를 침해할 수 있다. 사회 전체 총합을 기준으로 하는 경우에는 100 명에게 10 만큼의 쾌락을 주고 10 명에게 90 만큼의 고통을 주는 행위도 좋은 행위가 된다. 이를 AI 알고리즘을 통해 일반화, 자동화시킴으로써 총합을 구성하는 계산식의 세부 구성 요소를 살피지 않고 총합이 양수인 경우를 모두 승인한다면, 소수 의견이 데이터화되어 반영되더라도 소수의 의견은 지속해서 소외될 것이다. 예를 들어, 장애인 전용 택시 한 대를 운영하는 비용과 일반 택시 다섯 대를 운영하는 비용이 같다면, 쾌락의 총합만 고려하는 AI 는 후자를 택하게 될 것이고 이는 우리 사회의 일반적인 정의 관념에 비추어 합리적이라고 보기 어려울 것이다.

벤담과 달리, 밀(John stuart Mill)은 쾌락의 질적 차이를 인정하였으며 질은 항상 양을 압도하므로 고급 쾌락은 항상 저급 쾌락을 넘어서서 선택된다고 하였다[15]. 밀은 쾌락의 질적 차이의 판단과 관련하여 실천적 반성(자기반성과 토론)을 기준으로 제시하였으며, 이러한 과정에서 다른 사람보다 좀 더 나은 판단자들로 드러난 '능숙한 판단자의 선호'가 쾌락의 질적 우월성을 결정한다고 주장했다[16]. 능숙한 판단자는 비교가 될 두 쾌락 모두를 경험해 본 사람들이자, 두 쾌락 모두를 평가할 수 있는 기질을 가지고 있고, 강한 성격을 소유한 적당하게

[15] "만약 두 가지 쾌락 중, 이 둘 모두를 충분하게 경험해본 그런 사람들에 의해서, 하나가 다른 것보다 훨씬 중시된다면, 설령 이러한 선호가 엄청난 양의 불만족을 수반한다는 것을 안다고 해도, 그들은 그것을 선호할 것이다, 그리고 쾌락의 많은 양을 선택할 수 있는 그들의 기질(nature)에도 불구하고 그것을 포기하지 않을 것이다. [그러면] 우리가 적은 정도의 가치를 표현하는 더 큰 양에 비하여 그 선호된 즐거움에 질적 우월성을 부여하는 것은 당연하다" J. S. Mill, Utilitarianism, 279 쪽

[16] "두 쾌락 중 어떤 것이 최고의 가치를 갖는지의 질문에 대해 ... 양쪽 모두에 대해 정통하고 있다고 인정되는 사람들의 판단은 최종적인 것으로 인정되어야만 한다. 양쪽에 모두 익숙한 사람의 생각이 외에 어떤 방법이 있겠는가?" J. S. Mill, Utilitarianism, 282 쪽.

잘 교육받고 지성을 갖춘 사람들이며, 또한 느낌과 의식을 가진 관찰력이 예리하고 자의식이 강한 사람들이다[17](김은미, 2015).

밀은 쾌락의 질적 차이의 의미나 질적 차이를 판단하는 방법, 질적 차이와 양적 차이의 비교 등에 대해 구체적으로 설명하지 않음으로써 비판을 받았는데(Guy Fletcher, 2008; Michael Hauskeller, 2011), 이러한 논의는 AI 에도 그대로 적용될 수 있다(김은미, 2015). 먼저 유능한 판단자들이 항상 합리적 결정을 하는지와 개인의 선택에 사회가 어디까지 관여할 수 있는지 등이 불분명한 상황에서, AI 가 유능한 판단자의 역할을 할 수 있는가라는 비판이 가능하다. AI 가 유능한 판단자가 되기 위해서는 고급 능력을 갖추고 모든 종류의 쾌락, 적어도 결과 도출을 위한 비교 대상이 되는 쾌락을 직접 경험해볼 수 있어야 하겠지만, 이는 불가하다. 또한 동일한 문제에 대해서 다양한 알고리즘의 AI 가 같은 결론을 도출함으로써 질적 우월성의 객관성을 입증할 수 있을지가 불분명하며, AI 의 결정이 대상이 되는 사람들로부터 정당하다고 인정될 가능성이 낮다는 맹점이 있다.

또한, 쾌락과 고통을 계산해서 행복을 극대화하기 위해서는 질적 차이가 계량화 되어야 계산에 반영할 수 있다. 그러나 예를 들어 시를 쓰는 행위가 놀이기구를 타는 행위보다 질적으로 우월하다고 하더라도 구체적으로 얼마나 차이가 있으며, 벤담이 양적 계산에서 활용한 척도인 강도나 지속성 등에는 어떠한 영향을 미치는지 알 수 없다. 더욱이 밀은 양적 부족에도 불구하고 질적으로 높은 쾌락이 선택된다고 하였는데, 고급 쾌락과 저급 쾌락의 질적 차이가 미미할 경우 양적 차이가 우선하는 것이 더 합리적인 상황이 생길 수 있다. 예를 들어 질적 수준이 5 이면서 10 만큼의 양적 쾌락이 있는 행위와, 질적 수준이 4 이면서 100 만큼의 양적 쾌락이 있는 행위에서도 언제나 전자가 우선한다고 볼 수 있을 것인가? 혹은 사회 전체적인 관점에서 대다수의 쾌락과 소수의 쾌락이 부딪히는 상황에서 소수의 쾌락이 질적으로 조금 더 높을 때, 어떤 것을 우선적으로 선택하게 될 것인지가 모호하다. 이와 같은 불확실성은 정량화되는 데이터를 기반으로 운영되는 AI 운영체계에서 활용되기에는 부적합할 것이다.

나아가, 밀은 벤담과 달리 즉각적인 공리가 아니라 장기적 관점에서의 공리 극대화를 주장(Sandel, 2009)하면서 벤담의 공리주의가 가진 난점을 극복하고자 하였다. 그러나 어느 정도의 범위를 단기와 장기의 기준으로 정할 것인지가 불분명하며, 과거 데이터의 학습에 기반하는 AI 가 장기적 관점의 공리를 예측할 수 있는 능력을 갖출 수 없다는 한계가 있다. 예를 들어, 휘발유와 경유 자동차가 발명되었을 때 AI 가 미래의 석유 고갈 및 환경오염 문제를 정확히 예측할 수 있었으리라고 기대할 수 없다.

결론적으로, '최대다수의 최대행복'을 주장하는 공리주의는 공통적으로 행복이 무엇인지는 정답이 없으며, 행복이 곧 정의이며 도덕적인 것이라고 도출될 수는 없다는 비판을 받는다. 또한 인간 행동에 대해 획일적으로 가치를 평가하는 것은 인권 침해이며, 일부 사람의 희생으로 다른 사람들이 이득을 받는 것을 허용한다는 문제에 직면한다(한희원, 2018). AI 의 특성을 고려할 때, 학습된 데이터에 의존하는 AI 가 과거와 전혀 다른 상황을 맞닥뜨리거나 미래를 예측해야 할 경우 쾌락의 질과 양을 계산할 수 없는 동시에, AI 의 블랙박스 효과로 인해 쾌락을 계산하는 알고리즘을 이해하고 관리 감독하는 데에 한계가 있다는 문제가 도출된다.

---

[17] 밀은 이러한 능력을 인간 고유의 고급 능력으로 간주한다. "보기 위한 관찰 능력, 예측하기 위한 추론과 판단, 결정을 하기 위해 자료를 수집하기 위한 활동능력, 결정을 하기 위한 판별능력, 무언가를 결정했을 때, 자신의 숙고된 결정을 지지하기 위하여 확고 부동함과 자기 통제의 능력" J. S. Mill, On Liberty, 85 쪽.

ii. AI 시대의 정의 – 자유주의적 평등

롤즈는 기본적 권리와 자유를 최대한의 유용성을 위한 도구로 전락시키는 공리주의의 결함을 비판하며, 이에 대한 대안으로 공정으로서의 정의(justice as fairness)을 주장했다. 롤즈는 사회 구성원 개개인의 자유를 인정하고 존중하면서도 복지와 같은 사회의 혜택을 제대로 받지 못하는 사람들에게 차별이 최소화되는 세상을 제공하기 위해 평등적 정의론을 구축하고자 했다. 공동체적 삶에서 공정한 규칙을 만들기 위해서는 개인별 조건과 환경이 같아야 하므로, 롤즈는 원초 상태와 무지의 장막이라는 개념을 도입했다. 즉 자유롭고 평등한 개개인이 자유롭게 사회적 선택을 할 수 있는 객관적 환경(원초상태)에서, 천부적인 제반 능력을 포함하여 스스로에 대한 그 어느 것도 전혀 알지 못하는 주관적 상태(무지의 장막)로 공동체의 삶의 원칙을 함께 결정한다고 가정한다. 롤즈는 정의의 원칙으로 기본권 평등 및 차등의 원칙을 제시했으며, 이러한 공정한 절차에 따라 도출된 결론이 정의로운 것이라고 하였다.

롤즈는 사람들이 특정한 개인을 희생하면서까지 사회의 행복을 극대화하려고 하기보다는, 오히려 모든 사람은 최소한 일정한 기본권(자유)은 모두가 절대적으로 평등하게 누려야 한다는 원칙에 합의할 것이라고 주장한다. 여기서 롤즈의 제 1 원칙인 기본권 평등의 원칙이 도출된다. 또한 사람들은 경제적 및 사회적 문제에 대해서는 일정한 정도의 불평등을 용인하여, 사회의 최소 수혜계층에게 이익이 돌아갈 때는 불평등을 인정할 것이라는 데에서 제 2 원칙인 차등의 원칙이 도출된다. 예컨대 변호사나 의사에게는 버스 운전사보다 더 높은 보수를 주는 식으로 불평등을 인정하면서도, 그들에게서 세금을 더 받아 빈곤층의 의료혜택을 늘리는 등의 환경 개선을 할 수 있다. 롤즈는 제 1 원칙은 제 2 의 원칙에 의해서도 훼손되어서는 안 되는 정의의 기본적인 원리라고 역설한다. 자유는 인간에게 너무나 기본적이고 소중한 것이어서 경제적인 이유나 사회적인 이득을 위해서가 아니라 오로지 자유 그 자체를 위해서만 제한될 수 있다는 것이다(한희원, 2018).

하지만, 롤즈는 수혜층의 기준, 분배의 정당성 및 불명확성 측면에서 비판을 받는다. 먼저 롤즈는 소득을 기준으로 가진 자와 수혜층을 구분할 수 있다고 하였으나, 소득이라는 단순한 잣대로 차별의 문제를 모두 설명할 수는 없다는 점에서 수혜 대상의 정의에 대한 고려가 부족했다는 비판을 받는다. 이를 데이터를 기준으로 움직이는 AI 시대에 적용하면, 수혜층에 대한 판단이 더욱 모호하고 편향적이게 될 위험이 있다. 예를 들어 근로 소득을 기준으로 할 때 무직으로 소득은 없으나 본인이나 가족의 자산이 충분히 많아 구직할 필요 자체가 없는 경우나, 노후 보장에 대한 준비가 다 된 상태로 은퇴한 노인층, 혹은 데이터 추적이 되지 않는 현금 등의 방법으로 소득을 얻는 경우가 모두 수혜 대상으로 분류될 수 있다.

또한 사회적으로 기회가 불평등하게 주어진 것 외에 인간의 타고난 지능과 같은 능력이나, 후천적 노력으로 성취한 덕성 등도 분배가 정당한 것인지, 분배를 한다면 얼마나 하는 것이 적절한지에 대한 이견이 존재한다는 현실적인 문제가 있다. 이에 따라 각 국가는 시대 및 환경적 배경 변화에 따라 복지제도에 대한 결정을 수정하고 있다. 이렇듯 사회적 합의가 부재하고 분배의 기준을 일반화해서 설명하기 어려운 상황에서 고정된 데이터를 기반으로 움직여야 하는 AI 알고리즘을 운영해야 한다면, 가변적인 정의에 대한 개념을 바탕으로 운영될 수 없다는 한계가 지적될 수 있다.

iii. AI 시대의 정의 – 공동체주의

마이클 샌델(Michael Sandel)은 공리주의, 자유주의, 공동체주의 등 정의에 대한 이전 연구들을 비판하면서 아리스토텔리스에서 유래한 공화주의 전통을 기반으로 한 '공공철학으로서의 정의론'을 제시한다. 공동체의 일원로서의 개인이 공동체의 좋은 삶에 우선권을 부여해야 하며, 이를 위해서는 개인 차원뿐만 아니라 국가 차원의 역할이 융화적으로 이루어져야 한다는 것이다. 샌델은 정의에 대한 정의와 정의를 수호하는 방식의 두 가지 차원에서 기존 연구를 비판한다. 먼저 샌델에 따르면 기존 연구들은 정의에 대한 기준의 가변성과 공동체에 대한 논의가 결여되었다는 점에서 비판 가능하다. 공리주의자들의 정의는 개인의 선택을 기반으로 일치된 기준을 통해 규정하는데, 이에 대해 롤즈는 좋음이나 선보다 옳음이 우선이라고 주장하며, 공정과 같은 단일기준을 내세워 공공생활에서의 개인의 특수성을 획일화할 위험이 있다고 지적한다. 또한 샌델은 기존 공동체의 가치를 추구하는 공동체주의자들에 대해서, 공동체주의는 공동선을 위한 것이기 때문에 기존의 가치와 달라질 수도 있고, 개선할 수도 있는 것이라고 주장한다.

샌델의 정의론은 올바름, 즉 공정이 아니라 좋은 삶에 우선권을 부여하여야 한다는 태도가 드러난다. 샌델은 서로 다름을 인정하고 공동체의 좋은 삶을 조화롭게 영위하기 위해서는 형식적 기회균등만으로는 불충분하다고 주장한다. 예컨대 단거리 세계 최고 기록 보유자인 우사인 볼트와의 100m 경쟁에서 아무리 출발 선상에서 기회가 균등하게 주어진다고 해서 그 결과를 공정하다고 말할 수는 없다고 말한다. 또한, 정의를 수호하는 방식에 있어 국가와 정치의 적극적인 역할을 강조하며 기존 연구를 비판한다. 자유주의자들은 국가가 삶의 가치문제에 있어 중립적이어야 한다고 주장하지만, 샌델은 국가를 포함한 공동체는 결코 가치문제에 중립적인 방관자가 되어서는 안 되고, 올바른 자율성과 품성을 갖춘 시민이 육성될 수 있도록 오히려 적극적으로 나서야 하고 행동해야 한다고 주장한다.

롤즈는 결과의 불공정성을 고려해, 복지를 통해 시작점을 최대한 유사하게 가져가야 한다고 주장하지만 샌델은 그러한 노력이 사회의 문화를 바꾸거나 공동체 구성원들의 공공선을 추구하는 태도를 이끌어내지 못하므로 부정적으로 본다. 예를 들어, 산업화와 급속한 경제 발전의 과정에서 장착된 능력주의 문화가 민주주의 제도의 발전과 결합하면서 공정한 절차가 정의로운 결과를 낳지 못하는 왜곡이 발생하였으므로, 샌델은 이러한 문화를 바꾸기 위해 사회적 연대를 회복해야 한다고 이야기한다. 우리의 삶은 수많은 다른 사람들에게 의존하며 살아가므로, 보상 체계는 공동체에 대한 기여의 관점에서 이루어져야 한다. 예를 들어 팬데믹 상황에서 간호사와 택배 노동자의 도움이 보상으로 이어질 수 있어야 한다. 또한 성공한 자들과 패배한 자들이 자신의 성공과 패배에 깃든 운의 역할을 살펴보아야 한다. 승자는 감사를 배우고, 패자는 자신의 굴욕을 당연시할 필요가 없는 문화가 필요하다[18].

샌델은 "공동체적 관점에서 공동선을 추구하는 좋은 삶"을 정의로 규명하지만, 공동체의 관점은 누가, 어떻게 결정할 수 있는지에 대해서는 이야기하지 않는다. 결국 다수가 공동체의 관점을 결정하게 되는 것이다. 예를 들어서, 샌델은 미국의 교육제도 중 SAT 제도를 비판하며, 시험보다는 고교 생활에 더 중점을 둬야 한다고 주장한다[19]. 그런데 교육의 목적이라는 공공선에서 생각할 때, 어떤 것이 옳은 것인가? 공동체적 시각에서 보면 SAT 시험 성적은 가계

[18] 『한겨레』 (2021.10.13) "센델 "능력주의 오만, 공동선에 대한 책임 망각하게 해".
[19] 『매일경제』 (2021.01.01) "[단독] 마이클 샌델 "능력•학력 중시가 계층 이동제약… 불공정 만들어"".

소득과 비례하므로 경제적 배경이 높은 성적을 결정하지만, 개인주의적 시각에서 보면 좋은 유전자에 따라 지능이 높아서 높은 성적을 얻었다고 이야기할 수 있다. 이 때 좋은 학교를 가게 소득에 따른 실력은 제외하고, 고등학교 성실성에 부합해서 보내주는 것이 교육의 목적에 부합하는 공공선적인 선택인가? 아니면 사회에 진출했을 때 더 많은 공익을 창출할 수 있는 성과를 가진 인재들을 더 배출하기 위해서 더 우수한 학생들이 더 좋은 대학교를 가도록 지원하는 것이 공공선적인 선택인가? 무엇이 옳은지는 결국 투표에서 이긴 쪽이 결정하게 될 것이라는 점에서, 샌델은 롤즈를 비판하면서도 개선책은 내놓지 않고 공리주의와 다름없는 결과를 낳는다는 문제가 있다.

또한 공동선의 정의에 있어서 시간과 범위가 모호하다는 문제가 있다. 쾌락에 대해서 질과 장기성을 중요시한 밀과 같이 범위나 우선순위, 구체적인 기준 등이 결여되어 있다. 위의 사례를 가져오자면, 공동선이 공익의 총합이라고 할 때 장기적 관점에서는 성적이 높은 똑똑한 학생이 좋은 대학을 졸업해서 더 좋은 상품이나 서비스를 제공하는 것이 사회적으로 더 이롭다고 볼 수 있다. 그러나 단기적 관점에서는 소득이 낮은 학생들은 좋은 대학에 입학할 수 없다는 점에서, 그 시점의 공익의 총합을 보면 다수가 불공정하다고 느낄 수 있기 때문에 공동선이 추구되지 않은 것이라고 할 수 있을 것이다.

이를 AI 와 연관지어 생각해보면, AI 는 스스로 가치 판단을 하는 것이 아니라 무비판적인 알고리즘에 따라 아웃풋을 산출한다는 특징이 있다. 따라서 AI 는 AI 를 설계하는 각 공동체의 가치관과 이익을 대표하는 제품이나 서비스를 도출하게 될 것이다. 위의 사례를 가져오자면, 고등학교에서의 성실성을 강조하는 공동체에서는 각 학년별 성취도를 자세히 평가할 수 있는 AI 평가 시스템을 개발할 것이고, 우수한 성적을 강조하는 공동체에서는 최종 시험인 수능 시험의 영역별 점수 비중을 높인 AI 평가 시스템을 개발하려 할 것이다. 이렇게 공동체마다 추구하는 방향성이 합의되지 않은 상황에서는 AI 가 도출하는 제품이나 서비스가 해당 공동체의 공공선을 증진시키는 데 기여하더라도, 그것이 다른 공동체의 공공선을 감소시키는 등 공공선의 간극과 충돌이 발생할 수 있다. 따라서 공익을 총합하는 전체 사회의 관점에서 본다면 공공선을 추구하지 않는 것과 다름없게 된다.

이와 같은 논의를 바탕으로 AI 시대에서 기존의 정의에 대한 논의의 부합성을 논하자면, 본 연구는 AI 알고리즘의 한계를 보완할 수 있는 방향성을 제시하기 위한 사회적 합의로서 롤즈식의 정의론이 적합하다고 판단한다. 먼저, 공리주의적 AI 는 쾌락과 고통을 정밀하게 계산하는 알고리즘을 만들기 어려우며 여러 가지 형태의 쾌락의 우월성을 비교할 기준이 모호하고, 최대다수를 강조함에 따라 데이터가 반영되지 않거나 적은 소수자의 권리가 배제된다는 한계가 있다. 물론 샌델식의 정의는 기존의 개인을 중심으로 한 정의에 대한 논의의 한계를 개선시킨다는 의의가 있지만, 그가 주장하는 공동체의 규모와 범위가 모호한 상황에서 결국 각 공동체가 추구하는 이익을 보장해주는 AI 를 제공한다면 그 공동체가 갖는 한계점을 AI 가 그대로 답습할 수밖에 없다. 즉, 샌델의 정의론은 공동체 및 공동선의 범위를 어떻게 정의하는지에 따라 무엇이 '정의'인지가 달라지며, 세부 공동체별 가치관 차이로 인해 전체 사회에 바람직하지 않은 결과를 낳을 수 있다는 난점을 갖는다. 공동체의 규모가 작을 때에는 전체 공동체의 이익에 반하는 방향성을 제시할 것이고, 반면 공동체의 규모가 클 때에는 거기에 속하는 못하는 이들이 소외되는 공리주의적 문제가 다시 발생할 수밖에 없다. 물론, 롤즈의 정의론 또한 분배의 대상이 되는 최소 수혜층의 기준이 모호하고, 정당한 분배를 위한 구체적인

기준이나 방법에 대한 설명이 결여되어 있다는 문제가 있다. 하지만, AI 알고리즘이 보유하고 있는 데이터 입력에서의 편향성 및 과거지향성이나 그로 인한 결과값의 편향성이라는 문제를 보완할 수 있는 것은 롤즈식의 정의적 관점이라는 것을 알 수 있다. 다시 말해, 이렇듯 모든 정의론에 결점이 있는 상황에서 앞서 언급한 AI 알고리즘에 내재된 문제점인 데이터의 과거지향성, 근접성, 편향성과 블랙박스 효과 등을 고려한다면, AI 알고리즘이 가진 편향성을 지적하고, 그것을 제도를 통해 개선할 수 있는 명확한 가이드라인을 설계한다는 점에서 롤즈의 방향성을 추구하는 것이 적절할 것이다. 롤즈식 정의를 바탕으로 한 AI 알고리즘에 대한 개선 방향성의 논의가 진전될 때만이 AI가 현실적으로 가변적인 사회경제적 상황 등을 고려해서 결정하도록 하며, 사회적 소외계층의 정의와 이에 대한 차등 임금 등의 구체적인 기준 마련에 대한 고민이 가능할 것이다. 다만, 이와 같은 사회적 협의를 위해서는 AI 데이터와 결론을 조정하여 AI가 편향된 결과만을 산출하지 않도록 하되 공동선을 추구하는 동시에 공동체적 협력을 통해 공동선을 도출하고자 하는 샌델의 공동체주의적 관점에 대한 수용 또한 고려되어야 할 것이다.

IV. AI 시대의 정의

이에 따라, 본 연구는 AI 시대의 정의를 구축하기 위한 방안으로 AI 운영 체계에 대한 사전적 보완 및 사후적 제재를 제시하고자 한다. 다만, 해당 논의를 진전하기 이전에 이와 같은 논의가 필요한 이유에 대한 이해가 필요하다. 다시 말해, AI 시대의 정의 또한 기존의 도덕적 신념이 변화해온 것과 같이 지속적으로 변화할 수 있다는 특징을 보유하는데, 이전 시대와 AI 시대의 정의에 대한 논의의 중요성이 상이하게 인식되어야 하는 이유는 두 가지의 상황이 사회 구성원들의 동의를 이끌어낼 수 있는지에 대한 기본 논의의 시작점 자체에 차이가 있다는 점에 있다. 예컨대, 지금까지의 사회 구성원들인 사람들은 이전의 공공적 차원의 도덕이라는 다소 추상적인 개념에 대해서 개인적 도덕적 신념의 변화에도 불구하고 원칙적으로 받아들일 수 있는 사회적 맥락에서 살아왔다. 대체로 공공의 이익을 추구하는 공적인 차원의 '사회 정의(social justice)'는 국가 또는 공공 기관을 기반으로 한 법적, 규칙적 체계를 통해서 오랜 시간 동안 추구되어 왔으며, 이는 사회를 유지 및 관리하기 위한 최소한의 정당한 원칙이라는 암묵적인 합의가 있었다. 그러나, AI 시대에 AI의 알고리즘을 기반으로 도출되는 결정들에 대해 과연 사회 구성원들이 정당한 결정이라는 것을 받아들일 수 있을 것인가에 대한 고민이 필요하다. 즉, 공공적 차원의 정당성을 보유한 것은 각 개인 간의 차이와 불일치에도 불구하고, 어떠한 공통의 관점을 참조함으로써 각자에게 받아들여질 수 있어야 하는 것(Quang, 2013)[20]이라는 점에서 AI 시대의 정의란 기술의 통합과 배치가 공공의 측면에서 정당화되어야 한다[21].

AI 시대가 도래하는 현재, AI 는 사회 구조의 기본을 구축하는 구성 요소의 일부가 되어가고 있으며, 이와 같은 상황에서 AI 의 설계는 개인들의 권리를 보장하고, 공동체적인 정의의 개념을 지지하는 실질적 속성을 구현해야 한다(Gabriel, 2022). 이에 따라, AI 시대에

[20] Quong, J. (2013) "Public Reason," *The Stanford Encyclopedia of Philosophy*.
[21] Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & Technology*, 31(4), 543-556.

적합한 정의에 대한 논의를 수립하기 위해서는 AI 운영 체계 수립의 과정에서 롤즈식의 분배적 개념을 포함한 정의의 원칙에 대한 고려가 필수적이다[22]. 다만, 이 과정에서 간과되어서는 안 되는 또 다른 핵심 내용은 AI 운영체계의 선택에 따라 데이터 값을 입력하고, 도출된 결과값 중 필요한 부분이 도출된다는 점에서 각각의 AI 의 특징은 해당 프로그램을 구축하는 각 객체의 개수만큼 다양하게 나타날 수 있다는 점이다. 이는 곧 AI 가 활용되는 영역과 방법이 다양해지는 만큼 각 AI 만의 특징이 상이하게 나타남으로써, 인간의 다양성만큼이나 서로 다른 특징을 보유한 서로 다른 AI 운영 체계가 발생하게 된다는 것을 의미한다. 그에 따라, AI 시대의 특징인 가치 판단이 결여된 알고리즘이 자동화되어 움직인다는 점을 제외하고 본다면, 이는 인간사회와 유사성을 띈다는 것을 알 수 있다. 그러므로, 결국 AI 시대에 정의를 구축하기 위해서는 인간 세계에서의 최소한의 정의로움을 구축하기 위한 제도적인 틀인 법적 체계와 유사한 틀을 제시해주되, 향후 AI 의 성장 방향성을 현재의 상황에서 가늠할 수 없기 때문에 할 수 있는 것을 명시하는 포지티브 규제의 특징을 띄는 것이 아닌, 해서는 안 될 것을 명시하는 네거티브 규제의 특징 또는 지지해야 할 방향성을 제시하는 가이드라인적 형식의 윤리규범이 필요하다. 하지만, AI 시대에서도 AI 운영 체계의 잘못된 결과 및 결과값에 따라 정의롭지 않거나 정당하지 않은 서비스 및 제품이 도출되어 사회적 정의에 영향을 미친다면, 그 경우에는 법적 체계에서의 제재와 같은 규제 및 제재가 발현되어야 한다.

그렇다면, 롤즈를 기반으로 한 AI 시대의 정의는 어떻게 규명되어야 하는가? 이에 대한 제안을 하기 위해서는 AI 서비스의 개괄적인 구조와 해당 구조 내에서 정의의 개념을 어떻게 적용시킬 수 있을지에 대한 고민이 필요하다. 왜냐하면 AI 는 자동화된 운영 체계임에 따라 정의로운 결정을 도출하기 위한 영향이 반영될 수 있는 영역 자체가 한정적이기 때문이다. 다시 말해, 앞에서 바라본 AI 의 입력값, 데이터베이스, 레이블링, 알고리즘, 결과값으로 이어지는 운영 체계와 그것의 문제점을 고려해보았을 때, 데이터베이스에서부터 알고리즘에 이르는 모든 단계는 자동화된다는 것을 알 수 있다. 그러므로, 롤즈의 정의가 적용되기 위해서는 원자료가 입력되는 입력값 부분과 알고리즘을 통해 도출된 결과값에 대해 영향을 미치는 방안을 고려할 수밖에 없을 것이다. 결론적으로, AI 시대의 정의를 위해서는 알고리즘의 데이터베이스를 구축하는 원자료에 대한 편향성을 사전적으로 보완하기 위한 보완책을 제시하고, 자동화된 알고리즘에서 정당성을 보유하지 않은 결과값이 도출된 후에는 그것에 대한 규제 및 제재를 통해 결과값을 사후에 보완하기 위해 노력함으로써, 소외되는 사람들을 줄여나가기 위해 롤즈식의 정의로움을 AI 시대에 구축하는 노력이 제공해야 한다.

i. AI 시대의 정의로움 구축을 위한 방안 – 사전적 보완(입력값)

먼저, 입력값을 의미하는 원자료 데이터에 대한 문제점을 데이터의 과거지향성, 근접성, 그리고 편향성이라고 지적하였는데, 이는 알고리즘을 구성하는 데이터베이스 자체가 편향성을 띌 수밖에 없게 만들고, 이를 보완하기 위해서는 롤즈식의 정의의 개념을 활용한 보완책이 필요하다. 특히, AI 시대에서 원자료 입력 데이터에 포함되지 못하는 이들을 어떻게 포함할 것인지에 대한 고민이 필수적이다. 왜냐하면 이들이 바로 롤즈의 시각에서 제시되는 무지의

[22] Hoffmann, A. L. (2017). Beyond distributions and primary goods: Assessing applications of rawls in information science and technology literature since 1990. *Journal of the Association for Information Science and Technology*, 68(7), 1601-1618.

베일의 상태에서 고려했을 때, 사회가 권리를 보장하는 영역 밖으로 내몰리는 소외자의 영역에 포함되는 이들이기 때문이다. 이 경우 우리는 원자료를 보완할 수 있는 방법을 제시함으로써, 정보화 시대의 선두에서 시대를 이끄는 서비스를 제공하는 AI 알고리즘에서 도태되어 AI 가 제공하는 서비스의 혜택을 받지 못하거나, 서비스 제공의 대상자로서 인정받지 못하는 사람들이 포함될 수 있도록 해야 한다.

이에 따라, 본 연구는 정의롭고, 정당성을 보유한 입력값을 바탕으로 한 데이터베이스를 확보하기 위한 노력으로써, 정부를 필두로 한 데이터 수집 및 운영에 대한 가이드라인 제시 및 직접적인 보완 데이터의 제공 등을 핵심 해결 방안으로 제시한다. 이는 곧 기존 아날로그 세계에서의 법치 기준(Završnik, 2020)을 의미하는 AI 에 대한 인권 준수 및 나아가 데이터베이스를 책임 있게 만들기 위한 AI 정의를 위한 체계 설계를 의미한다. 이와 같이 사전적 보완책에 집중하는 이유는 입력값의 편향성을 완화시키기 위해 직접 규제 및 제재가 어려우며, 그에 따라 간접적인 제안 및 보완 데이터의 제공 등을 통한 보완에 집중해야 하기 때문이다. 그 이유는 AI 의 입력값은 대체로 확보 가능한 데이터를 통해 운영되는 경향이 존재하며, 입력값을 제재한다는 것은 현실적으로 가용하기 어려운 데이터를 알고리즘에 입력하지 못했다는 현실적인 어려움에 대한 고려가 없는 것으로 간주할 수 있기 때문이라는 데에 있다.

그렇다면, 이에 대한 방안은 정부를 중심으로 이루어지며, 크게 AI 에 대한 윤리 가이드라인을 제시하는 것과 실질적인 보완적 데이터를 제공하는 것으로 분류된다. 먼저, AI 윤리 관련 가이드라인은 EU, 한국 등 이미 다수의 국가에서 제시하고 있다. 예를 들어 2019 년 유럽연합 집행위원회 EU 나 OECD 도 AI 윤리 권고안을 마련했고, 민간기업인 구글이나 마이크로소프트도 자체적인 AI 윤리 원칙을 마련해서 준수하고 있으며, 국내에서도 과학기술정보통신부, 방송통신위원회가 나름의 가이드라인을 제시(박종선, 2017)한 바 있다. 하지만, 위의 논의의 한계는 추상적 수준에서의 논의에 머물러 있다는 것이며, 이를 개선한 사례로는 뉴욕 시의회에서 통과시킨 알고리즘적 의사결정 투명성에 대한 법률[23] (Završnik, 2020) 정도이다. 이 법에 따라 지방 자치 단체가 사용하는 알고리즘의 공정성과 타당성을 감시하기 위한 태스크포스가 설치되었다. 하지만, 자동화되는 알고리즘에 직접적인 영향을 미치기 어려우며, 나아가 양질의 데이터, 즉 데이터 입력 차원에서 소외되는 사람들의 데이터를 어떻게 입력 및 보완할 것인지에 대한 논의가 제시되지 않는다는 점에서 그 한계를 지적할 수 있다.

이를 보완하기 위해서는 정부가 직접적으로 데이터 입력값에서 제외되는 데이터를 선제적으로 확보하여 제공하는 방법에 대해 고민해볼 수 있다. 예를 들어, 다수의 사회적 이슈에 대해 사회적 정의의 차원에서 논의할 때, 다수에 속하지 못해 소외되는 인구층인 아동, 노인, 장애인층 등과 같은 경우에는 온라인 접근성이 기타 인구층과 대비하여 봤을 때 다소 낮게 나타날 수 있다. 이 때에 정부는 그들과 관련된 데이터가 AI 알고리즘 데이터베이스에 입력값으로 제공될 수 있도록 직접적인 오프라인 인터뷰, 설문조사 수행 및 기확보하고 있는 정부 데이터의 일부를 통한 자료 재가공 등의 노력을 통해 자동화된 데이터 입력에서 누락되는 자료에 대한 보완데이터를 제공할 수 있다. 이와 더불어, 국가는 AI 관련 가이드라인을 통해 정부가 제공하는 보완적 성격의 데이터를 AI 운영체계를 활용한 서비스, 제품 및 의사결정을

[23] Vacca, J., & Rosenthal, H. (2018). *A Local Law in relation to automated decision systems used by agencies.* Technical Report. The New York City Council. .

추진할 때에 데이터 입력값의 일부로써 활용하도록 해당 사안을 강력하게 권고함으로써 입력값 편향성을 제한하기 위한 최소한의 노력을 수행해야 한다.

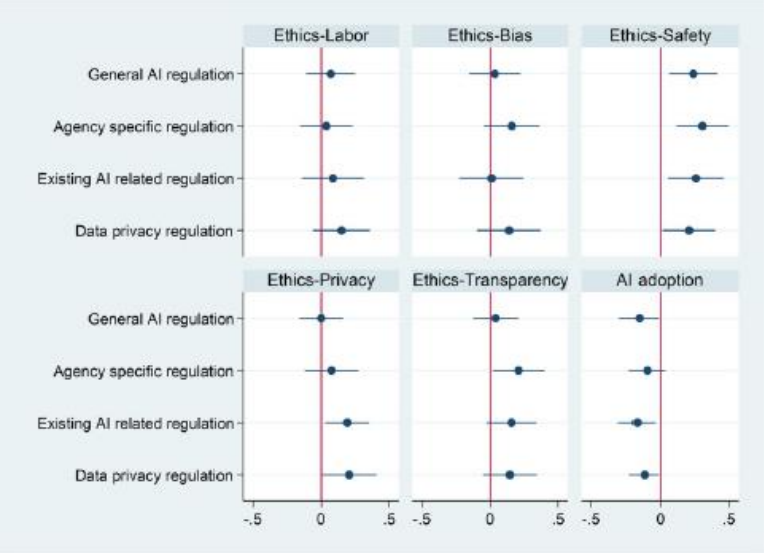
구체적으로 말하자면, 정부는 AI 가 가장 많이 활용되는 서비스 영역을 도출한 다음 해당 영역에서 소외되는 사용자들에 대한 심층 인터뷰를 통해 원자료 데이터를 구성하고, 이를 AI 를 활용한 서비스와 제품을 구상하는 모든 층위의 이용자에게 활용할 수 있도록 지속 권고해야 하며, 데이터가 부재할 경우 그에 대한 보완적 데이터를 구축할 수 있는 체계를 수립해야 한다. 이에 대한 필요성은 비단 경영자뿐 아니라, 정부 서비스 구축 차원에서도 핵심적으로 논의되어야 할 부분이다. 예를 들어, AI 서비스를 활용하여 야간 택시/대중교통 운영 전략을 구축한다고 하였을 때, 기존 AI 운영 체계가 확보할 수 있는 원자료 데이터, 즉 입력 가능 값은, 온라인에 대한 접근성이 높은 20-50 대층의 교통 활용 패턴에 치우쳐 있을 것이다. 만약 해당 데이터만을 기반으로 한 교통 서비스를 회사 및 정부에서 수립하게 된다면, 해당 입력값에 포함되지 못하는 아동, 노인, 장애인, 외국인 등의 데이터가 누락된 상황에서 서비스가 도출될 것이며, 이는 롤즈식의 정의로운 AI 시스템이라고 볼 수 없다. 그러므로, 이 경우에는 정부가 해당 소외계층의 택시 및 대중교통 운영 현황을 조사하고, 관련된 데이터를 온라인에 게시하는 것을 넘어, 해당 서비스를 구축하는 사기업에도 선제적으로 보완 데이터를 활용할 수 있도록 권고하는 역할을 수행함으로써 편향되기 쉬운 AI 입력값에 대한 보완 방법의 권고 및 해결 방안을 제시할 수 있을 것이다.

실제로, 사전적 보완책을 제공했을 때에 기대되는 효능은, 잠재적인 인공지능 규제가 AI 와 관련된 윤리적 이슈에 대한 관리자들의 인식과 AI 기술을 채택하려는 그들의 의향에 미치는 영향에 대한 연구(Cuéllar et al., 2022) 사례를 통해 알 수 있다. 연구에서는 미국의 1,245 명의 관리자들에게 무작위로 온라인 설문조사를 진행하였다. AI 기술이 점점 더 상용화되고 있으나 아직 잠재적 가치를 실현하기 위해서는 핵심 관행을 개발해야 하며, 알고리즘의 투명성, 데이터의 편향성, 안전과 사생활에 대한 우려 등을 고려해 어떻게 AI 를 비즈니스 전략에 접목시킬 수 있을지 고민해야 한다는 글을 제시하면서 통제집단에게는 규제를 특별히 언급하지 않았으나, 실험 집단은 네 개의 그룹으로 나누어 각각 규제에 대한 정보를 글에 포함하였다. 예를 들어, 일반적인 AI 규제 정보를 노출시킨 그룹 1 에게는 기존에는 연방이나 주 정부가 AI 를 특별히 규제하지 않았으나, 2020 년 새로운 알고리즘 책임법의 등장으로 기업들은 AI 시스템 디자인, 수집된 데이터 등 AI 시스템의 사용 현황을 공개해야 할 것이라는 내용을 추가하였다. 산업 특화 규제 정보를 노출시킨 그룹 2 에게는 2020 년 FDA 가 보건 산업 기업의 AI 제품이 시판되기 전에 성능을 검토하고 알고리즘 수정사항을 승인하도록 하는 새로운 규제 프레임워크를 제안했다는 내용, NHTSA(National Highway Traffic Safety Administration)가 자율 주행 기술 혁신을 저해하지 않도록 규제보다는 자발적 지침을 내리는 방향으로 규제 장벽을 제거할 것이라는 내용, 공정위가 AI 기업들의 마이크로 타겟팅, 제품 추천 엔진 등에 대한 감시를 강화하면서 소비자 보호를 위한 청문회를 열었다는 내용을 추가하였다. 기존 법적 규제를 노출시킨 그룹 3 에게는 기존의 불법행위나 노동 관련 법률이 AI 기업에도 적용되고 있으며 앞으로도 지속될 것이라는 내용을 추가하였고, 개인 정보 보호 규제를 노출시킨 그룹 4 에게는 2020 년 발효되는 CCPA(California Consumer Privacy Act)에 따라 캘리포니아 기업들은 개인 데이터의 사용 및 저장 현황과 개인 정보 보호 규칙을 어떻게 준수하는지 공개해야 한다는 내용을 추가하였다. 이후 모든 참여자들에게 AI 기술 채택, 예산 할당, AI 관련 혁신, 윤리적 이슈, 노동 등 다섯 가지의 주제와 관련된 질문을 하였다.

그 결과 AI 규제에 대한 정보는 AI 와 관련된 안전(safety), 사생활(privacy), 편향성(bias/discrimination), 투명성(transparency) 윤리 이슈의 중요성에 대한 관리자들의 인식을 제고하는 것으로 드러났다. 특히 관리자들은 전반적으로 AI 시스템이 해를 끼칠 경우 정량화 및 측정될 수 있는 구체적 사례를 연상하기 쉬운 안전과 같은 윤리적 영역에 대해서는 더 민감하게 반응하는 반면, 광범위한 해결책을 찾기 어려운 차별이나 투명성 이슈에 대해서는 덜 호의적으로 응답하였다. 그리고 기존 정보 규제나 개인 정보 보호 규제를 접한 관리자들은 사생활과 데이터 보안에 대한 인식이 상당히 제고되는 경향을 보였으며, 산업 특화 규제를 접한 관리자들은 편향성과 투명성에 대한 지각이 높아졌다. 한편, AI 규제 정보는 AI 기술 채택에 대한 관리자들의 의향을 감소시키는 것으로 나타났으며, 나아가 관리자들로 하여금 직원들의 AI 교육이나 AI 소프트웨어 패키지 구매와 같이 AI 채택을 위해 투자하기보다는 윤리적 이슈들을 포함한 AI 전략을 세우는 데에 비용을 사용할 의사를 증가시켰다.

이를 통해 AI 이용에 대한 규제가 존재한다는 것을 인지하는 것만으로도 AI 를 활용하는 이들의 의사결정이 달라질 수 있다는 것을 알 수 있다. 다시 말해, 만약 정부가 AI 의 입력값과 관련된 가이드라인을 마련하고, 해당 내용을 AI 알고리즘을 구축하는 민간 개발자 및 개발사에게 전달한다면 그 내용이 직접적인 규제가 아니더라도 정부의 가이드라인에 가까운 AI 입력값을 활용하기 위해 노력하게 될 것이다. 결국, 정부가 모든 AI 입력값에 대해 명확한 방향성을 제시하지 못하더라도 데이터 활용자들이 입력값을 확보하고 알고리즘을 구축하는 과정에서 정부가 제시하는 정의로운 데이터 확보 및 활용 방향성에 대해 인지하게 하는 것만으로도 최소한의 사회적 피해, 즉 단순 데이터 입력을 통한 소외의 극대화라는 문제를 피할 수 있을 것이다.

[그림 3] 윤리적 이슈 및 AI 도입에 미치는 AI 규제의 영향력 관련 계수도[24]



[24] 본 그래프는 Cueller et al. 2022 에서 추출했으며, 점은 회귀 분석의 계수 추정치를 나타내고, 막대는 95% 신뢰 구간을 나타내며, 각 계수 추정치는 각 처리 그룹과 관리 그룹 간의 차이를 나타냄

ii. AI 시대의 정의로움 구축을 위한 방안 – 사후적 규제 및 제재(결과값)

나아가, 위와 같은 불완전한 데이터를 기반으로 한 알고리즘은 일방향성이라는 특징에 기반하여, 데이터 불완전성을 심화시키기 때문에 시간이 소요될수록 그 결과 또한 편향될 가능성이 높아진다. 또한 알고리즘은 자동화되어 운영된다는 특징을 고려했을 때, 알고리즘 자체에 정의로운 결과값이 나오도록 개선할 수 없기 때문에 정부는 이에 대한 차선책으로 결과값에 대한 대안을 제시해야 한다. 다시 말해, 결과값이 정의롭지 못한 경우 정부는 그에 대한 사후의 제재 또는 규제를 통해 보완하기 위해 노력해야 한다. AI 시대가 도래함에 따라 자동화된 알고리즘을 통해 도출된 결과값이 사회 정의에 부합될 수 있도록 정부는 인공지능 개발자, 정부, 법률가, 철학자 등 전문가와 함께 AI 가 안전하게 활용될 수 있도록 환경을 정비하고, 다양한 문제를 해결하기 위해 개발과 활용 및 응용 과정에 대한 표준 및 관행을 구체화해야 한다(송기복, 2020). 왜냐하면 AI 운영 체계를 통해 도출된 결과값이 잘못된 의견을 도출하고, 이를 통해 사회경제적으로 중요한 의사결정이 이루어질 경우, 인간들의 의사결정 논의과정과 달리 결과값에 대한 추가 논의와 비판 없이 추진될 가능성이라는 위험성을 보유하기 때문이다.

이 경우, AI 를 활용한 결과값 도출 이후에 AI 시대의 정의로운 결정을 위해 제시할 수 있는 방안은 사후적 규제 및 제재일 것이며, 결과값의 활용 방안에 대해 그에 대한 범위와 방향성이 상이하게 도출되어야 한다. 예컨대, AI 알고리즘을 통해 도출된 결과값이 정의로운 값을 보유하지 못하고 있을 때에 그 활용법은 해당 결과를 바탕으로 의사결정을 위한 논의가 이루어지는 것과 직접적으로 그것을 바탕으로 의사결정이 이루어질 때 등 두 가지로 분류된다. 이 중에서 사회정의적 차원에서 더욱 문제가 되는 경우는 결정값에 대한 인간의 관여도가 낮은 후자의 방법이라고 할 수 있다. 그렇다면, 두 가지의 서로 다른 AI 결과값 활용 방법에 따라 AI 시대의 사회적 정의 수호를 위해서는 어떠한 차원의 상이한 규제 및 제재가 이루어져야 하는가? 이에 대한 답변을 제시하기 위해서는 AI 결과값이 서비스 및 상품을 위한 어떠한 수단(method)으로 활용되는 영역에 대한 고려가 필요하다. 왜냐하면 AI 결과값이 미칠 수 있는 영향력의 범위 및 규모에 따라 그에 대한 제재의 방법과 강도가 달라져야 하기 때문이다.

AI 결과값이 수단으로 활용되는 방법으로는 크게 단순 데이터 제공, 상품 및 서비스 제공, 의사결정 자문 제공 등으로 분류 가능하다. 해당 분류의 특징은 단순 데이터 제공으로부터 의사결정 자문 제공으로 갈수록 AI 결과값이 사회적으로 갖는 영향력의 크기가 커진다는 것이다. 예를 들어, 단순 데이터 제공의 경우에는 AI 알고리즘을 활용하여 도출된 값을 심화된 서비스 제공을 위해 제공하는 것을 의미하며, 그 예로는 한강 대교의 CCTV 분석을 통한 높은 자살 가능성을 보유한 사람이 등장했을 때 위기 경보를 제공하는 것이 있다. 이 경우 AI 를 통한 판단은 심화된 수준의 가치판단이 필요한 영역이라기보다 단순히 관련된 정보를 분석하여 발생 가능한 사고에 대한 예측치를 제공하는 수준에 머무르게 된다. 이에 반해 상품 및 서비스를 제공하는 AI 와 같은 경우에는 AI 결과값을 중심으로 도출된 예측 가능한 정보를 바탕으로 실생활에서 활용 가능한 상품 및 서비스를 직접적으로 제시한다는 차이가 존재한다. AI 를 통해 확보된 개인별 활동 내역에 대한 분석을 바탕으로 새로운 서비스 및 상품을 제공하는 예로는 개인적 차원에서는 관심사를 고려한 맞춤 도서, 최적화된 목적지 주행 방법 제시 등의 서비스와 개인별 상황에 맞는 옷, 도시락, 의료기기 등의 상품이 있을 것이다. 하지만 거시적인 측면에서 보면, 교통, 교육, 복지, 의료 등 실생활 밀접 영역에서의 서비스들 또한 포함되기 때문에 만약 AI 데이터 분석만을 통해 서비스가 제공되었을 때 상대적으로 AI 입력데이터에서 누락될 가능성이 높은 인구층은 실생활에서 누릴 수 있는 서비스 영역이 현격히 축소될 수 있다는

문제점이 더욱 심화될 수 있다. 그럼에도 불구하고, 이와 같은 서비스들은 의사결정 자문 제공과 달리 상대적으로 해당 서비스가 사회적으로 미치는 영향력의 범위가 좁고, 규모가 작다. 반면, AI 의 결과값이 의사결정 자문을 제공하는 경우는 정부, 규제당국, 정책결정자 및 법원 등 정치·사회·경제 등 주요 영역에서의 중요성을 가지는 의사결정에 영향을 미치는 사례를 의미한다. 이전 사례들과 달리 이 경우에는 AI 의 결과값이 사람들의 삶에 직접적으로 영향을 미칠 수 있는 영향력을 보유하기 때문에 그에 대한 중요도가 상당히 높다고 볼 수 있다.

그렇다면, AI 운영 체계를 활용하여 도출된 값을 활용하는 방법에 따른 정부의 규제 및 제재의 범위의 정도는 어떻게 달라져야 하는가? 결과값의 영향력이 확대될수록 그에 대한 규제가 강화되어야 하며, 나아가 그 범위 또한 더욱 광범위해져야 한다. 예컨대, AI 체계를 활용한 결과값이 단순 데이터 제공에 그칠 때에는 정부가 해당 결과값의 편향성이 높을 경우를 정의하고, 해당 결과가 사회에 직접적인 악영향을 미치지 않도록 그것을 보완해야 한다는 취지의 권고 가이드라인을 제시하고 후속 조치를 지속적으로 진행해야 한다. 반면, 서비스 및 상품 제공의 경우에는 다수에 포함되지 않은 소외자들에 대해서는 맞지 않는 상품과 서비스가 제공되거나, 제공되지 않을 수 있다. 이와 같은 경우에 대비하여 정부는 소외자들을 위한 상품과 서비스를 함께 제공하도록 규제해야 할 수 있다. 예를 들어, AI 운영 체계를 통해 분석한 데이터를 기반으로 줄기세포 치료법을 제공한다고 하였을 때, 해당 서비스의 이익을 공유할 수 있도록 소외자들에게도 유사한 서비스가 제공될 수 있도록 정정해야 한다. 이에 반해, 만약 해당 서비스가 인종 차별을 의미하는 슬로건이 적힌 티셔츠를 판매했을 때에는 정부가 적극적 개입을 통해 소수자의 인권 보호를 위해 해당 상품 판매를 금지해야 한다.

AI 가 수집 가능한 데이터에서 누락될 가능성이 높은 소비자층에 대한 고려가 부재한 AI 결과값을 바탕으로 제시되는 서비스 및 상품 제공의 문제점에 대한 규제 필요성은 실질적으로 나타날 수 있는 예시들을 통해 검증 가능하다. 이에, 본 연구는 일상생활에 중요한 민간과 정부가 모두 서비스를 제공하는 사회인프라 영역인 교통 영역에 대한 사례 분석을 통해 무비판적인 AI 분석 자료 수용을 바탕으로 제공하는 서비스 및 상품 제공에 대한 규제 필요성을 제기하고자 한다. 이와 같은 문제점의 발생 가능성을 검증하기 위해 본 연구는 한국교통연구원에서 수행한 2020 년도 교통 물류 대국민조사 자료를 기반으로 상대적으로 AI 시대에 소외될 수 있는 소비자군인 노년층이 직면할 수 있는 문제를 알아보았다. 이를 위해 높은 연령대일수록 교통물류적 상황에 대해 어떻게 느끼는지를 도출해봄으로써 현재 그들의 직면하고 있는 문제는 어떠한 것이 있는지를 제시한다. 이와 같은 분석은 AI 가 모집하는 원자료에서 소외될 가능성이 높은 연령대의 문제점이 어떻게 심화될 수 있는지에 대한 예측을 가능하게 한다. 이를 위해 핵심 종속변수를 연령대로, 핵심 독립변수로는 대도시 거주 여부, 교통수단의 요금이 비싸다고 평가하는지 여부(고속도로, 대중교통, 택시, 고속버스 통행료 등), 정부의 교통정책의 방향성(서비스 개선 필요성) 등의 변수들을 포함하였다[25].

본 연구의 분석 결과에 따르면, 거주지, 대중교통의 교통수단에 대한 평가, 정부의 교통정책에서의 서비스 개선 필요성 등의 변수가 통계적으로 유의미하게 나타났다. 다시 말해,

[25] 한국교통연구원은 2020년 3월 18일(수) ~ 4월 17일(금)의 기간동안 전국(제주 제외)에 거주하는 만 20~79세 남녀를 대상으로 각 5천 명씩 2회에 걸쳐 설문조사를 수행하였다. 본 연구에서는 한국교통연구원으로부터 교통 현안 대국민 조사의 응답 데이터 원본을 제공받았으며, 그 중 비용, 시간, 복지 등의 주제를 포함하는 설문조사 부분을 활용하여 회귀분석을 실시하였다.

서울 및 경기권 등을 포함하는 대도시에 살지 않고, 대중교통 및 고속버스 비용이 비싸다고 생각할수록, 택시 비용은 비싸다고 생각하지 않을수록 연령대가 높은 것으로 나타났다. 또한 정부가 교통정책을 개선하기 위해 교통서비스 비용을 현행 수준으로 유지하되 서비스를 개선해야 한다고 생각할수록 연령대가 높아졌다. 이를 통해 교통편의 시설이 부족한 지역에 거주하거나, 일상 생활에서 필수적인 대중교통 또는 저렴한 비용으로 장거리를 이동할 수 있는 교통수단인 고속버스 비용에 대해 부담을 느낄수록 연령대가 높다는 것을 알 수 있다. 즉, 현재의 상황에서 생활에 필수적인 양질의 대중 교통 서비스를 만끽하고 있지 못할수록 연령대가 높다는 것을 의미한다. 이와 같은 상황에서 AI 를 통한 분석만을 바탕으로 새로운 교통 서비스나 상품을 제시하게 된다면, 인구 비중이 높은 노년층의 교통 시장에서 느끼는 불편함을 개선하지 못하는 문제가 확대될 수 있다. 왜냐하면 노년층의 경우 인구가 집중되어 있는 대도시에 주거하고 있지도 않으며, 대중교통 및 고속버스 비용에 대해 부담을 느낄 만큼 경제적으로도 소외되어 있을 가능성이 높았기에 AI 분석을 통해 새로운 상품 및 서비스를 제안하는 과정의 우선순위에서 밀려날 수밖에 없을 것이다. 그렇다면 결과적으로 현재에도 상대적으로 소외되어 있는 노년층이 AI 시대가 도래할수록 그들이 필요로 느끼는 대중교통의 개선은 AI 분석을 통해서는 해결되기 어렵다는 것을 알 수 있다.

[표1] 교통환경 요인 평가와 연령 간의 상관 관계

		연령		
		계수	표준편차	P값
사회인구 통계학적 변수	성별	0.0449	0.0463	0.332
	대도시 거주	-0.1496	0.0529	0.005**
교통환경 요인 평가	고속도로 비용	-0.0414	0.0903	0.646
	대중교통 비용	0.2867	0.1047	0.006**
	택시 비용	-0.1693	0.0821	0.039*
	고속버스 비용	0.3331	0.1460	0.023*
	교통편의 서비스 개선 필요성	0.2235	0.0487	0.000***
절편		4.3034	0.0795	0.000***
N		5,000		
R-squared		0.0091		
Adj R-square		0.0077		

*** p<0.001, ** p<0.01, * p<0.05

이에 반해, AI 의 결과값이 중요 의사결정에 대한 자문을 주는 경우에는 정부가 보다 적극적으로 그 역할의 범위를 한정시키고, 규제해야 한다. 이에 대한 논의는 이미 사법부 결정에 대해 AI 의 역할 증대에 따라 나타날 수 있는 문제점을 통해 진행되고 있다. 예를 들어, 유럽 국가들은 알고리즘이 기본적인 자유에 미치는 영향과 '알고리즘을 책임 있게 만드는 방법(algorithms accountable)'에 대한 우려를 공유하고, 이에 따라 유럽사법효율화 위원회(CEPEJ: European Commission for the Efficiency of Justice)는 2018 년 말 '사법시스템에서 AI 사용에 관한 유럽헌장'을 채택하여 사법분야에서 언급된 위험을 완화하고자 하였다[26]. 다만, 이와 같은 노력을 통해 AI 의 정당하지 못한 결정에 대한 제재가 보다 직접적으로 실행되어야 하는 경우도 있다는 것을 인지하고, 대응할 수 있어야 한다. 예컨대, AI 를 활용하여 정부의 정책 결정을 내린다는 방안을 고려 중이라면, 정부는 AI 의 결정을 최소 인원의 관련 부처 담당자들을 통한 논의를 거친 후에 활용하도록 해야 한다는 규제 방침을 제시하거나, AI 결과값을 활용하여 도출된 제안은 적어도 최소 숫자의 서로 다른 데이터 베이스를 이용하여 도출된 AI 결과값들을 함께 비교 분석하여 최선의 대안책을 찾아서 선정해야 한다는 방식의 정형화된 규제 프로세스를 제시하고, 해당 체계가 확고히 운영될 수 있도록 규제를 통해 강제해야 한다.

사후적 규제에 대한 필요성은 2017 년 인공지능 분야 전문가들을 대상으로 수행된 설문조사 결과를 통해서도 알 수 있다. 본 연구는 전문가 집단을 대상으로 한 설문조사 데이터[27]를 통해 AI 결과물 규제 필요성에 영향을 미치는 요인이 무엇인지 회귀분석을 검증했다. 이를 위해 'AI 결과물에 대한 규제 및 제재 필요성'을 종속변수로 설정하고, 핵심 독립변수로는 AI 의 위험성(일상적 업무의 자동화, AI 위험 발생 가능성 등), AI 잠재적 위험의 책임 소재(정부, 제조사, 소비자 등), AI 위험 발생 시 필요 정책(수익자 비용 부담 정책) 등을 포함했다[28].

---

[26] EUROPEIA, U. European Ethical Charter on the Use of Artificial Intelligence. *Judicial Systems and their environment. Strasbourg*, 3-4. c

[27] 한국행정연구원은 2017 년 6 월 5 일~8 월 4 일의 기간동안 전국 235 명의 인공지능, 신기술, 기술위험관리 분야에 종사하는 학계(연구기관), 정부, 기업 전문가를 대상으로 설문조사를 수행하였으며, 그 중 본 연구에서는 인공지능으로부터 기대되는 편익과 위험성, AI 잠재적 위험의 책임 소재, AI 위험 발생 시 필요 정책 주제를 포함하는 설문조사 부분을 활용하여 회귀분석을 실시하였다.

[28] 각 주요 변수들에 대한 설문지 질문은 아래와 같으며, 각 문항은 5 점 척도로 응답되었다:

1) 종속변수: "인공지능 위험에 대비하여 인공지능 위험비용·책임체계 논의 및 마련이 얼마나 중요하다고 생각하십니까?:
2) 독립변수-AI 로부터 기대되는 편익과 위험성:
- 일상적 업무의 자동화: 인공지능으로부터 얻을 수 있는 다음 편익들을 얼마나 얻을 수 있다고 생각하십니까?
- AI 위험 발생 가능성: "'인공지능은 위험 발생 가능성이 높다'는 의견에 대해 얼마나 동의하십니까?"
- AI 위험성이 자녀층에 미칠 영향: "'인공지능 위험은 내 아들과 딸들에게 영향을 미칠 것이다'라는 의견에 대해 얼마나 동의하십니까?"
- AI 의 평등의 가치 침해 가능성: "인공지능 위험으로 인해 평등의 가치는 얼마나 침해받을 것이라고 생각하십니까?"
3) 독립변수-AI 잠재적 위험의 책임소재:
- A 정부/민간개발사(자)/소비자: 인공지능의 잠재적 위험에 대해 다음 주체들이 평가할 책임이 얼마나 있다고 생각하십니까?
4) AI 위험 발생시 필요 정책:
- 수익자가 비용을 부담하는 정책: 인공지능 위험에 대비해 인공지능 위험이 발생할 경우 해당 수익자가 그 비용을 부담할 수 있는 정책 (위험증명의 수익자 부담, 수익자 사전 재정보증 제도 등)의 정책이 얼마나 필요하다고 생각하십니까?"

본 연구의 분석 결과에 따르면, ‘일상적 업무의 자동화’, ‘AI 위험 발생 가능성’, ‘AI 위험성이 자녀층에 미칠 영향’, ‘AI 위험이 평등의 가치를 침해할 가능성’, ‘정부 책임’, ‘수익자가 비용을 부담하는 정책’ 등의 변수가 통계적으로 유의미하게 나타났다. 다시 말해, 전문가들은 AI 로부터 기대되는 편익 및 위험성에 대한 변수들 중에서는 AI 로 인해 일상적 업무가 자동화될 것이고, AI 의 위험 발생 가능성이 높으며, 그것이 평등의 가치를 침해할 것이라고 생각할수록, 그러나 해당 AI 위험이 자녀층에게 영향을 미치지 않을 것이라고 생각할수록, AI 규제에 대한 필요성을 높게 평가하였다. 나아가, AI 의 잠재적 책임에 있어서는 정부에게 책임이 있다고 평가하며, AI 위험이 발생할 경우 해당 수익자가 그 비용을 부담할 수 있는 정책이 필요하다고 생각할수록 AI 규제 마련의 필요성이 높다고 응답했다. 이를 통해 AI 로 인해 업무의 자동화 대체와 일상의 변화가 상당할 것으로 예상되고, AI 의 높은 위험성이 특정

[표 2] AI 결과물 규제 필요성에 영향을 미치는 요인

		AI 결과물에 대한 규제 및 제재 필요성		
		계수	표준편차	P값
사회인구 통계학적 변수	남성	-0.0346	0.1057	0.744
	연령	-0.0478	0.0805	0.553
	교육수준	-0.0044	0.3407	0.897
	근속 연수	0.0718	0.0519	0.168
AI로부터 기대되는 편익 및 위험성	일상적 업무의 자동화	0.1329	0.0664	0.047*
	AI 위험 발생 가능성	-0.1321	0.0624	0.035*
	AI 위험성이 자녀층에 미칠 영향	0.1527	0.0679	0.026*
	AI의 평등의 가치 침해 가능성	0.1726	0.0484	0.000***
AI 잠재적 위험의 책임 소재	정부	0.2550	0.0645	0.000***
	민간개발사(자)	-0.0366	0.0641	0.568
	소비자	0.0459	0.0453	0.313
AI 위험 발생시 필요 정책	수익자가 비용을 부담하는 정책	0.3816	0.0575	0.000***
절편		0.1476	0.4412	0.738
N		235		
R-squared		0.4130		
Adj R-square		0.3813		

*** p<0.001, ** p<0.01, * p<0.05

계층에게만 이득이 되고 사회적 약자층에게는 더 불리하게 작용하여 불평등이 확산될 것이라고 판단할수록 AI 결과물에 대한 제재가 필요하다고 여긴다는 것을 알 수 있다. 또한 AI 의 위험이 자녀층에게 미치지 않을 것이라고 생각한 것에서 미루어 볼 때 이러한 전문가들은 AI 결과물에 대한 규제를 통해 AI 위험을 대비 및 조정함으로써 미래에 미칠 위험을 줄일 수 있을 것이라고 여기고 있으며, 정부 주도 하에 AI 의 수익자가 위험에 대한 책임을 지는 방향, 즉 불평등의 위험을 감소시키는 방향으로 규제해야 한다고 생각한다는 것을 알 수 있다.

결과적으로, 이를 통해 전문가들은 AI 의 위험성이 크다고 느끼지만, 정부가 그에 대한 책임 소재를 명확히 하고 배상 및 책임에 대한 규제와 제재를 제시한다면 장기적으로는 그 위험성을 극복해 나갈 수 있다고 판단하고 있다는 것을 알 수 있다. 다시 말해, 이는 현재의 AI 시대는 문제점을 도출할 가능성이 높고, 그에 대해 정부의 적극적인 규제 역할의 중요도가 높아지고 있다. 나아가, 이는 정부가 사전적 보완책을 제시하는 것을 넘어, 사후적 결과물이 적절한 방향으로 도출되고 사회 정의를 추구하는 방향으로 활용될 수 있도록 가이드라인을 제시해야 한다는 것을 의미한다. 이와 더불어, 사회적 정의를 수호하는 사후적 결과물 규제를 위해서는 규제 및 제재 대상이 되는 대상을 구체적으로 지정하고, 각각에 결과물 대해 발생 가능한 문제점에 대해 비용 부담뿐만 아닌, 법적 제재에까지 이르는 폭넓은 차원의 규제 정책이 구체적인 정책 방침을 통해 제시되어야 할 것이다.

V. 결론

AI 의 영향력이 증대되고 있는 현재, 우리 사회는 더욱 빠르고 효율적으로 움직이게 되었고, 그에 따른 문제점 또한 우리 삶에 더욱 빠르고 깊게 침투하게 되었으며, 이에 따라 사회에서는 AI 에 대한 규제 방안을 적극으로 논의하고 있다. 반면, 그에 비해 AI 의 영향력이 확대될수록 어떠한 문제점이 도출되는 것뿐만 아니라, 이에 대한 해결책을 제시하기 위해 수반되어야 하는 사회적 차원의 합의를 이끌어 내기 위해서 어떠한 고민을 해야 하는지에 대한 근본적인 논의가 미흡한 상황이다. AI 는 알고리즘을 통해 인간의 뇌와 유사한 의사결정을 할 수 있지만, 인간과 달리 단기간에 인간이 감당할 수 없는 수준의 연산을 수행하고, 나아가 해당 연산을 끝없이 반복하면서 더 나은 방향을 찾아간다. 하지만, 만약 이와 같이 지속적으로 발전하는 AI 가 사회가 원하지 않는 방향으로 개발되면 그에 대한 피해는 해당 AI 를 사용하지 않은 다수의 대중에게까지 미칠 수 있다. 그러므로 AI 운영 체계 전반에 대한 이해를 바탕으로, 해당 체계의 주요 단계별로 어떠한 문제가 도출되는지에 따라 이를 해결하기 위해서 사회적 차원에서 수행되어야 할 논의가 무엇인지에 대해 고민이 필수적이다.

이에 따라, 본 연구는 먼저 AI 알고리즘 체계 운영과정에서 도출될 수 있는 입력값의 편향성과 과거지향성과 의심을 하지 않고 자동화되어 운영되는 알고리즘에 따라 AI 를 통해 제시되는 결과값이 사회적으로 긍정적이지 않은 결과를 도출할 수 있다는 점을 지적한다. 하지만, 여기서 더욱 중요한 것은 이와 같은 AI 활용에 따른 예측 가능한 문제점이 사회적으로 부정적 영향을 끼치지 않거나 적어도 최소화할 수 있도록 노력하는 방법에 대한 방향성을 제시하는 것이다. 이를 위해서 본 연구는 AI 는 사회적 공공선에 대한 가치관을 보유하고 비판적으로 사고하는 인간과 다르다는 점에 중점을 두고, AI 에서 활용되는 입력값이

사회적으로 합의된 방향에 근접하기 위해서는 정부 주도의 사전적 보완책을 제시해야 하고, 그 결과값 또한 사회적으로 유용한 결과값을 도출할 수 있도록 사후적 제재를 통해 규제 및 감시해야 한다고 주장한다.

하지만, AI 시대의 도래에 따른 올바른 대응을 위해서는 규제 및 제재의 방법에 대한 논의뿐만 아니라, 그와 같은 논의가 어떠한 사회적 방향성을 지녀야 AI 의 사회적 부작용을 완화시킬 수 있는지에 대한 고민이 필수적이다. 사회적 '정의'에 대한 논의는 해당 사회가 나아가야 할 방향성에 대한 사회적 합의를 도출한다는 점에서 역사적으로 의미를 보유해왔다. 인간 사회는 개인의 숫자만큼의 다양한 서로 다른 인격이 공동체를 이루고 살아가기 때문에 최소한으로 해당 사회를 지탱해줄 수 있는 사회적 가치에 대한 논의는 개인들의 사고방식뿐만 아니라 사회 전반의 운영 방향성에 영향을 준다는 점에서 의미가 있다. 왜냐하면, 해당 사회가 동의하는 가치가 무엇인지에 따라서 그 사회의 구성원들뿐만 아니라, 그 사회가 지향하는 방향성을 제시하는 역할을 수행할 수 있기 때문이다. 그러므로, 시대와 사회가 변화할수록 '정의'에 대한 논의의 방향성은 지속적으로 수정되어 왔다. 그렇다면, AI 시대에 필요한 정의는 무엇인가? AI 는 인간 사회에 영향을 미치는 수단이지만, 인간과 같은 감정과 비판적 사고를 보유하고 있지 못하기 때문에 AI 운영 체계 활용 그 자체만으로는 사회적 지지를 받지 못한다. 반면, AI 를 유익하게 활용할 수 있는, 즉 정의로운 사회를 만들기 위해 추구되어야 할 AI 의 활용 방향성을 제시할 수 있는 '정의'의 가치가 사회적으로 선행된다면, AI 알고리즘을 활용하여 도출된 결과가 사회적으로 정의로운 가치를 포함할 수 있을 것이다.

이에 대해 본 연구는 AI 시대에 가장 중요하게 고려되어야 할 문제점은 데이터의 편향성과 그로 인해 야기될 수 있는 사회 불평등적 가치를 포함한 결과값의 이용이라고 말한다. 자동화되어 운영되는 AI 운영 체계는 자동화 자체가 제공해줄 수 있는 효율성과 편의성을 보유하나, 이로 인해 해당 자동화된 알고리즘에서 도태되는 소외를 스스로 해결할 수 없음에 따른 사회적 불평등에 대한 위험성 또한 보유하고 있다. 예컨대 현재의 AI 알고리즘은 온라인을 중심으로 입력값을 자동적으로 확보하고, 그를 통해 도출한 결과값은 다시 기존 알고리즘을 강화하고 있으며, 이는 곧 온라인이라는 채널에서 정보를 제공할 수 있는 사회적 다수에 대한 서비스 및 상품을 고려하고 제공할 수 있다는 점에서 공리주의가 추구하는 최대 다수의 최대 행복을 추구할 수 있는 방법으로 인식될 수 있다. 하지만, 공리주의에 대한 주된 비판인 다수에 포함되지 못하는 소외된 인구들에 대한 고려가 이루어지지 않는다는 주장은 AI 시대에도 동일하게 제시될 수밖에 없다. 이와 더불어 AI 운영 체계는 해당 알고리즘을 구축하는 특정 집단 및 공동체의 가치관을 추구하는 방향성으로 활용되기 때문에 공동체주의가 추구하는 공동체적 시각에서 문제를 파악하고 그를 해결하기 위한 방향성을 제공할 수 있는 역할도 수행하고 있다. 그러나, 공동체별로 추구하는 가치관과 그에 대한 판단 기준이 없기 때문에 양극화가 심화되는 사회에서 문제점으로 지적하는 공동체 간의 충돌을 오히려 심화시킬 수 있다는 비판을 피하기 어려워보인다. 결국, AI 시대는 현재 대부분의 국가들이 자유주의를 추구하나 그 안에서 도태되고 소외시킬 수 있는 사람들을 위한 복지 제도를 제공하는 것과 같이 최소한의 사회적 안전망을 구축할 수 있는 롤즈식 자유주의적 평등주의를 추구해야 할 것이다. 즉, AI 알고리즘 구성 단계에서부터 입력값을 제공하고, 결과값을 보완하는 과정 전반에 롤즈식의 분배적 개념을 포함한 정의의 원칙에 대한 고려가 있을 때만이 AI 시대의 정의를 지켜나갈 수 있다. 이를 위해서는 알고리즘의 데이터베이스를 구축하는 원자료에 대한 편향성을 사전적으로 보완하기 위한 보완책을 제시하고, 자동화된 알고리즘에서 정당성을 보유하지 않은 결과값이 도출된

후에는 그것에 대한 규제 및 제재를 통해 결과값을 사후에 보완하기 위해 노력함으로써, 소외되는 사람들을 줄여나가기 위해 롤즈식의 정의로움을 AI 시대에 구축하는 노력이 제공되어야 한다.

AI 시대는 고도화된 기술력을 바탕으로 사회 전반에 영향을 미치고 있으며, 이로 인해 다수의 사람들의 삶은 더욱 윤택해지고 유익해질 수 있다. 하지만, 기술의 발전은 언제나 그로 인해 인간사회에서 또 다른 소외와 불평등을 야기시키는 결과를 불가피하게 제공하기도 한다. 이러한 사회적 문제를 최소화하기 위해서 사회적 정의에 대한 논의는 지속되어 왔으며, 정의에 대한 사회적 합의를 이끌어나가고 그것을 제도적으로 구체화하는 데에 정부의 영향이 현격한 중요성을 보유해왔다. AI 시대 역시 그 근본은 새로운 기술의 도입에 따른 사회적 구조의 변화, 그리고 그로 인해 발생하는 불평등의 심화라는 문제점의 발생이라는 다소 전통적인 차원의 논의 구조로 단순화될 수 있다. 하지만, 과거와 달리 한 번 야기된 불평등은 시스템화 및 자동화되어 움직이는 운영 체계 자체에서 제고될 수 있는 기회를 박탈당할 수 있다는 점에서 정부의 선제적이고 구체적인 대응에 대한 중요도가 더욱 높다고 생각된다. 그러므로, 세계적으로 AI 규제에 대한 논의의 필요성이 부각되기 시작한 지금이 정부가 AI 시대에 필요한 사회적 정의의 개념을 재수립하고, 정부가 사전적 데이터 보완과 사후적 결과값 규제라는 이원적 역할을 수행할 수 있는 구조를 수립하는 데에 있어 다시 오지 않을 골든타임이 될 것이다.

VI. 참고문헌

강준호. (2016). 공리주의적 전통과 벤담의 독창성. *범한철학*, 81, 87-121.

고선규. (2021). 인공지능 (AI) 과 정치의 관계 맺기: AI 는 통치수단일 수 있는가?. *정치와 공론 (구 정치와 평론)*, 28, 73-106.

구윤희. (2004). John Rawls 의 공리주의 비판과 정의론에 대한 연구.

김동현, & 이청호. (2022). 선이해와 데이터 편향-윤리적 인공지능의 도덕 온톨로지 구축을 위한 예비작업. *인공지능인문학연구*, 10, 155-176.

김원철. (2019). '최대다수의 최대행복'에 대한 계보학적 일 연구-프란시스 허치슨부터 제레미 벤담까지. *철학연구*, (60), 177-207.

김은미. (2017). 밀의 공리주의에 있어서 쾌락의 질적 차이. *철학탐구*, 47, 57-77.

김은미. (2015). 밀과 쾌락의 질적 테스트에 관한 해석. *철학탐구*, 40, 127-148.

김효은. (2021). 인공지능 편향식별의 공정성 기준과 완화. *한국심리학회지: 일반*, 40(4), 459-485.

박도현. (2022). 인간 편향성과 인공지능의 교차. *서울대학교 법학*, 63(1), 139-175.

박정기. (2010). 공리주의의 대안으로서 롤즈의 정의론. *동서사상*, 9, 275-296.

박종선. (2017). 인공지능에 대한 주요국의 대응전략 및 한국의 정치발전을 위한 제언. *법학논총*, 41(3), 35-73.

방준성, & 조수현. (2021). Ai 기술을 활용한 공공서비스 확대 방안. *KISDI AI Outlook*, *2021*(5), 5.

변순용. (2020). 데이터 윤리에서 인공지능 편향성 문제에 대한 연구. *윤리연구*, *1*(128), 143-158.

선종수. (2020). 의료 인공지능에 대한 형법적 고찰-왓슨 (Watson)을 중심으로. *법과 정책연구*, *20*(3), 249-274.

손권상, & 윤혜선. (2021). 키워드 네트워크와 BERT 모델을 활용한 인공지능 관련 국내외 법학연구 동향과 함의. *공법연구*, 50(1), 407-444.

송기복. (2020). 인공지능 (AI) 시대의 도래와 법제도의 방향에 관한 논의-독일의 인공지능 정책을 중심으로. *경찰법연구*, 18(2), 177-203.

심우민. (2018). 인공지능 기술과 IT 법체계: 법정보학적 함의를 중심으로. *동북아법연구*, 12(1), 55-86.

안현석. (2019). 허치슨의 도덕감과 도덕적 승인-[아름다움 및 덕 관념의 기원에 대한 탐구]를 중심으로. *범한철학*, 92, 87-112.

양삼석. (2007). 공리주의의 정치체계적 원용: 벤담의 의회입법을 중심으로. *철학논총*, 47, 159-181.

양천수. (2017). 인공지능과 법체계의 변화-형사사법을 예로 하여. *법철학연구*, *20*(2), 45-76.
유재흥, 추형석, & 강송희. (2021). 유럽(EU)의 인공지능 윤리 정책 현황과 시사점 : 원칙에서 실천으로. *SPRi 이슈리포트*, IS-114.
윤상필, 권헌영, & 김동욱. (2017). 건전한 인공지능 생태계 형성을 위한 규범적 전략과 법의 역할. *홍익법학*, *18*(2), 1-29.
윤혜선. (2021). 인공지능의 사회적 수용성과 법제도의 기능적 관계에 관한 소고. *경제규제와 법*, 14(1), 32-56.
정성훈. (2021). 인공지능의 편향과 계몽의 역설에 대한 반성적 접근. *철학연구*, *132*, 199-227.
정원섭. (2020). 인공지능 알고리즘의 편향성과 공정성. *인간· 환경· 미래*, (25), 55-73.
최경진. (2021). 인공지능의 사법적 쟁점. *저스티스*, (182-2), 151-171.
한지영. (2022). 인공지능과 법-인공지능 창작물의 권리귀속에 관한 검토. *아주법학*, 15(4), 335-370.
한희원 (2018). *인공지능(AI) 법과 공존윤리*. 서울: 박영사.
허유선. (2021). 인공지능 시스템의 다양성 논의, 그 의미와 확장-인공지능의 편향성에서 다양성까지. 철학· 사상· 문화, (35), 201-234.

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, 28(4), 399-416.
Agarwal, S., Krueger, G., Clark, J., Radford, A., Kim, J. W., & Brundage, M. (2021). Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818.*
Bentham, J. (1973). An Introduction to the Principles of Morals and Legislation. The Utilitarians. *Garden City and New York: Anchor Books*, 37.
Bentham, J. (1988). The Principles of Morals and Legislation / Jeremy Bentham. Buffalo, *N.Y. : Prometheus Books*. p. 23.
Bentham, J., Hart, H., & Burns, J. (1996). An introduction to the principles of morals and legislation (*The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*). Oxford: Clarendon. p.177
Binns, R. (2018). Algorithmic accountability and public reason. *Philosophy & technology*, 31(4), 543-556.
Campbell, R. W. (2020). Artificial intelligence in the courtroom: The delivery of justice in the age of machine learning. *Colo. Tech. LJ(18)*, 323.

Cuéllar, M. F., Larsen, B., Lee, Y. S., & Webb, M. (2022). Does Information About AI Regulation Change Manager Evaluation of Ethical Concerns and Intent to Adopt AI?. *The Journal of Law, Economics, and Organization*.

EUROPEIA, U. European Ethical Charter on the Use of Artificial Intelligence. *Judicial Systems and their environment. Strasbourg*, 3-4. C

Fletcher, G. (2008). The consistency of qualitative hedonism and the value of (at least some) malicious pleasures. *Utilitas*, *20*(4), 462-471.

Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, 151(2), 218-231.

Haakonssen, K. (1990). Natural law and moral realism: the Scottish synthesis.

Hauskeller, M. (2011). No philosophy for swine: John Stuart Mill on the quality of pleasures. *Utilitas*, 23(4), 428-446.

Hoffmann, A. L. (2017). Beyond distributions and primary goods: Assessing applications of rawls in information science and technology literature since 1990. *Journal of the Association for Information Science and Technology*, 68(7), 1601-1618.

Hutcheson, F. (1753). *An inquiry into the original of our ideas of beauty and virtue: in two treatises*. R. Ware., p.125

Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 3819-3828).

Le Bui, M., & Noble, S. U. (2020). We're missing a moral framework of justice in artificial intelligence. In *The Oxford Handbook of Ethics of AI* (p. 163). Oxford University Press.

Mill, J. S. (1859). Utilitarianism (1863). *Utilitarianism, Liberty, Representative Government*, 7-9.

Mill, J. S. (1975). On liberty (1859).

Otterbacher, J., Bates, J., & Clough, P. (2017, May). Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 6620-6631).

Quong, J. (2013) "Public Reason," *The Stanford Encyclopedia of Philosophy*.

Rawls, J. (2004). A theory of justice. In *Ethics* (pp. 229-234). Routledge.

Sandel, Michael J. (2009). *Justice : what's the right thing to do?*. New York : Farrar, Straus and Giroux

Sloan, R. H., & Warner, R. (2020). Beyond bias: Artificial intelligence and social justice. *Va. JL & Tech.*, 24, 1.

Ulloa, R., Richter, A. C., Makhortykh, M., Urman, A., & Kacperski, C. S. (2022). Representativeness and face-ism: Gender bias in image search. *New media & Society*, 14614448221100699.

Vacca, J., & Rosenthal, H. (2018). *A Local Law in relation to automated decision systems used by agencies*. Technical Report. The New York City Council.
Wolfe, R., & Caliskan, A. (2022). Markedness in Visual Semantic AI. *arXiv preprint arXiv:2205.11378*.
Završnik, A. (2020, March). Criminal justice, artificial intelligence systems, and human rights. In *ERA Forum* (Vol. 20, No. 4, pp. 567-583). Springer Berlin Heidelberg.

『경향신문』(2020.5.22) “(3)인공지능이 그린 그림, 예술인가 기술인가”.
『글로벌이코노믹』(2021.4.19) “AI, ‘의심’ 없어 인간 사고력과 차이 있다”.
『디지털투데이』 (2020.12.23) “과기정통부, 인공지능(AI) 윤리기준 마련”.
『리걸타임즈』 (2021.8.13) “[리걸타임즈 IT LAW] AI 규제 동향과 기업의 대응”.
『매일경제』(2021.1.1) “[단독] 마이클 샌델 "능력·학력 중시가 계층 이동제약…또다른 불공정 만들어"”.
『한겨레』(2021.10.13) “센델 “능력주의 오만, 공동선에 대한 책임 망각하게 해””.
『한국경제』 (2017.5.2) “’인공지능 법관’ 사람을 심판하다”.