과학철학

AI 윤리란 무엇인가?

2021년 5월 31일

이상욱

What is AI Ethics?

AI 윤리의 시대?

정말 인공지능AI이 우리 시대의 대세인 것 같다. SF 영화에나 등장하는, 사람과 구별되지 않는 안드로이드 로봇은 아직 먼 미래의 꿈이지만, 그보다는 훨씬 일상적이고 널리 퍼져 있는 AI가 우리에게 친숙한 휴대전화라는 기술적 대상 안에 내장되어 이미 일상생활 속 깊이 파고들고 있다. 이뿐만이 아니다. 오늘 저녁 어떤 영화를 볼 것인지를 결정하거나 휴가 때 읽을 책을 선택하는 과정에서도 많은 사람들이 이미 AI의 도움을 받고 있다. AI가 추천해 주는 선택지는 아직까지는 가끔 성가실 정도로 엉뚱한 것일 수도 있지만 상당히 많은 경우 꽤 쓸 만하다는 느낌이 들기도 한다. 처음에는 엉뚱해 보였던 추천 영화가 막상 보니 정말 내 취향에 딱 맞는다고 느낄 수도 있다. 이런 상황이면 조만간 기술이 더 발전해서 '나보다 나를 더 잘 아는' AI가 현실화되지 않을까 기대해 볼 수도 있다.[6]

AI의 일상화만큼이나 최근 국내외에서는 AI가 제기하는 여러 인문학적, 사회과학적 쟁점을 학술적으로, 실천적으로 탐색하는 연구 및 관련 활동도 활발하다. 전통적으로 인간만이 할 수 있었던 법률, 의료, 세무 등의 일자리 영역에서도 AI 활용이 늘어나면서 대량 실업 사태가 일어날 수 있다는 종말론적 두려움과 이를 정반대로 해석해서 인간이 노동으로부터 해방된 자유를 얻게 되리라는 유토피아적 기대가 함께 제시되고 있다. 어느 상황이 실현되든 하나의 해결책으로 논의되고 있는 '기본소득' 개념은 어느덧 상식적 담론이 되었다.[11]

AI에 대한 이런 다양한 쟁점을 통합적으로 다루는 분야를 최근 국제 논의 맥락에서는 대게 AI 윤리ethics라는 개념으로 포괄한다. 경제적으로 발전한 나라들의 모임인 OECD에서 AI의 개발이 가져다줄 혜택과 위험을 고려하여 사회적으로 수용가능한 수준의 절충점을 찾으려는 노력을 할 때도 AI 윤리 원칙ethical principles이라는 용어를 사용하고, 최근 유엔 기구 중에서 가장 활발하게 AI 윤리 논의와 규범적 틀 마련을 위해 노력하고 있는 유네스코도 AI 윤리라는 용어를 사용한다.[12,13] 이뿐만이 아니다. 세계적으로 가장 큰 전기전자공학자 단체인 IEEE는 AI가라는 단어가 줄 수 있는 불필요한 의인화 등을 걱정하여 AI라는 용어보다는 A/ISAutonomous Intelligent System이라는 용어를 선호한다. 그런 IEEE 또한 A/IS의 설계 단계에서부터 '윤리원칙에 일치하는 설계Ethically Aligned Design' 개념을 강조하며 아예 그와 관련된 국제 표준 마련에 힘을 쏟고 있다.[7]

그런데 국내에서는 AI 윤리라는 용어 자체에 대해 어색해하거나 불편해하는 사람들이 많이 있다. 이런 태도의 배경에는 윤리를 가치와 무관한 과학에 연결시키는 것이 부당하다는 직관이 있다고 볼 수도 있다. 자료를 조작하거나 다른 사람 연구를 표절하는 등 연구부정행위를 저지르지 않는 한 과학이나 기술은 윤리와는 무관한 가치중립적 영역이라는 생각일 것이다.[1]

바람직한 과학은 가치와 무관해야 된다는 생각은 여러 이유로 정당화되기 어렵다.[1] 하지만 이 주제에 대한 본격적 탐구는 다음 기회로 미루어두고 오늘은 일단 상당수의 사람들이 AI라는 기술적 대상에 대해 사람들의 개인적 행동에 적용되는 '윤리'라는 개념을 적용하는 것 자체가 이상하다고 느낀다는 점에서 출발해 보자. 이런 분들일수록 AI 윤리 논의 전체가 AI 관련 과학기술 연구의 '발목을 잡으려는' 비생산적 논의라고 규정하는 경향이 많다. 특히 과학기술 진흥과 연구개발의 효율성에 집중하는 정부 관료들 중에서는 국제적으로 이루어지는 AI 윤리 논의 자체가 형용모순이라고 생각하거나, 우리보다 AI 연구가 앞선 기술 선진국들이 윤리 논의로 우리나라와 같은 AI 기술 후발주자의 발을 묶으려는 '사다리 걷어차기' 전략이라고 의심하는 분조차 있다.

윤리倫理 vs. ethics

AI 윤리에 대해 국내외에서 왜 이런 차이가 발생하는 것일까? 여러 이유가 있겠지만 필자는 공약불가능한incommensurate 두 개념, 즉 윤리倫理와 ethics 사이의 의미 차이가 중요한 이유라고 생각한다. 일단 그 이야기부터 해보자.

"

AI 윤리라는 용어 자체에 대해 어색해하거나 불편해하는 사람들이 많이 있다.

이런 태도의 배경에는 윤리를 가치와 무관한 과학에 연결시키는 것이 부당 하다는 직관이 있다고 볼 수도 있다.

”

우리의 일상적인 언어 직관에 따르자면, ‘윤리’는 지극히 개인적인 사안에만 한정된다는 느낌이 있다. 이 직관은 표준국어대사전의 ‘윤리’에 대한 정의, “사람으로서 마땅히 행하거나 지켜야 할 도리”와도 일치한다. 이 정의에서 연상되는 상황은 천륜을 어기고 부모를 학대하는 행위나 상식적인 허용 범위를 넘어 극단적으로 자기이익만 챙기는 행위가 될 것 같다. 즉, 우리말에서 윤리란 개인이 누구에게나 명백하게 도리에 어긋나는 행동을 하는 것과 긴밀하게 관련되는 개념이다. 표준국어대사전은 윤리 개념의 용례로 채만식의 <낙조>라는 소설에 등장하는 “아내가 있는 사람이 한 다른 여자와 연애를 하고 어쩌고 한다는 것은, 나의 윤리로는 허락할 수 없는 패덕이었다.”는 문장을 들고 있는데, 이 문장에서도 우리의 윤리 개념이 개인적 사안과 관련된 것이며 명백한 잘잘못을 다룬다는 특징이 잘 드러나 있다.

이제 이런 윤리 개념으로 AI 윤리라는 표현을 살펴보면 누가 봐도 이상하다고 느끼지 않을 수 없다. 일단 AI 윤리에서 다루는 내용은 최근 문제가 된 AI 챗봇 이루다의 사례처럼 지극히 사회적이고 논쟁적이다. 이루다 사건이 문제라는 점에 대해서는 대체적으로 사회적 합의가 이루어졌지만, 다른 많은 AI 윤리 쟁점은 그렇지 않다. 예를 들어 AI 알고리즘의 투명성을 높이거나 설명가능성을 강하게 요구하다 보면 AI의 효율성이 저하되거나 민감 정보의 유출 가능성이 높아질 수도 있다.[9] 이처럼 현재 AI 윤리에서 논의되고 있는 내용은 (당연히 개인적 영역도 포함하지만) 많은 경우 사회적 수준에서 문제를 파악하고 해결책을 마련해야 하는 부분이고, 대부분의 경우 문제점 분석이나 해결책 마련 과정 자체가 많은 관련 집단의 이익과 다양한 가치를 종합적으로 고려해야하기 때문에 논쟁적이고 지난한 사회적 숙고를 요구한다.[5] 우리말의 ‘윤리’ 개념으로 AI 윤리를 제대로 이해하기 어려운 것도 무리는 아니다.

그럼 이제 영어의 ethics는 어떤 의미인지 살펴보자. 어원을 따져 보면 ethics는 고대 그리스어에서 ‘인격character’을 뜻하는 단어 ethos, 그리고 라틴어에서 ‘관습customs’을 뜻하는 단어 mores와 깊은 관련이 있다. mores라는 단어는 우리가 흔히 ‘도덕적’이라고 번역하는 영어 ‘moral’의 어원이기도 하다. 우리 일상표현에서도 윤리적과 도덕적을 서로 혼용해서 쓰듯이 영어에서도 (철학적으로 엄밀하게 구별할 때를 제외하면) 이 둘을 혼용해서 쓰는 경향이 있다. 그래서 옥스퍼드 영어사전에서 제시된 ethic의 정의는 “A set of moral principles, especially ones relating to or affirming a specified group, field, or form of conduct”이다. 이 정의에서 주목할 점은 ethic의 정의에 특정 집단, 분야, 행위의 종류가 등장한다는 사실이다. 이는 앞서 지적했듯이 ethic의 어원에 특정 집단이나 분야마다 공유되는 올바름의 기준이 다를 수 있는, ‘관습’의 의미가 포함되어 있다는 점과 일맥상통한다. 그리고 이러한 특징은 우리말의 ‘윤리’와 달리 영어의 ethic이 특정 개인의 행동 자체만이 아니라 그 행동의 사회적 의미까지를 본질적으로 포함하고 있음을 시사한다.

서양 문명의 기원이라고 여겨지는 그리스-로마 시대의 ethic에 해당하는 개념이 이처럼 개인적 수준과 사회적 수준을 가로지르고 있다는 사실을 염두에 두면, 황우석 연구팀의 논문조작 사건으로 촉발된 ‘연구 윤리research ethics’라는 개념이나 최근 강조되고 있는 ‘전문직 윤리professional ethics’ 개념이 결코 ethic 개념을 최근에 확장된 것이 아니라는 것을 짐작할 수 있다. 그보다는 이들 용어는 특정 집단에 고유한 내적 규범을 의미하는 ethic 본래의 의미에 충실한 것이

라는 사실이 자연스럽게 이해된다. 과학 연구자가 연구만 열심히 하면 되지 따로 윤리가 왜 필요하냐는 생각은 우리말의 '윤리' 직관을 따른다면 이해될 수 있는 반응이지만 영어의 ethic을 비롯한 국제적 기준에 따른다면 부적절한 반응이라고 볼 수 있는 것이다.

이 지점에서 오해의 여지를 제거할 필요가 있다. 필자는 우리말의 '윤리' 개념이 틀렸고 서양의 ethic 개념이 올바르다고 주장하는 것이 아니다. 그런 지적은 수number 개념으로 자연수는 틀린 개념이고 보다 포괄적인 정수나 실수 개념만이 진정한 수 개념이라고 주장하는 것만큼이나 터무니없다. 개념은 원칙적으로 맞고 틀리고의 문제라기보다는 정의의 문제이다. 필자의 지적은, 예를 들어 AI 윤리 관련 국제 논의에서 대부분의 나라는 모두 ethic의 의미를 배경으로 참여하는데 우리만 우리말에 고유한 '윤리' 개념을 갖고 참여한다면 생산적인 의사소통이나 논의 참여가 어려울 것이라는 점이다. AI ethics와 관련하여 국제적으로 통용될 수 있는 방안을 만들거나 법 제도화 등을 추진할 때 우리가 반드시 명심해야 할 부분이 바로 이것이다.

그렇다면 이렇게 개인과 사회를 가로지르는 의미의 윤리적 논의가 개인행동의 선악에 초점을 맞춘 우리의 윤리 논의와 구체적으로 어디에서 차이가 날까? 앞서 소개한 여러 AI 윤리 국제 논의에서 분명하게 부각되는 차이점은 우리가 사회적으로 추구해야 할 가치가 여럿이라는 사실, 그리고 그 가치들 사이에서는 종종 충돌이 일어난다는 사실이다. 이런 상황에서 공정하고 효율적인 윤리적 해결책은 거의 대부분의 경우 고려해야 할 여러 가치를 사회적으로 수용 가능한 방식으로 맞교환trade-off하는 방식으로 얻어지게 된다. 그리고 그런 과정에서 직관적으로 '좋은 것들' 사이에 절충이나 선택을 해야 하는 경우도 발생한다. 개인의 행동에 대한 선악 판단에서 암묵적으로 전제되는 '명백함'이나 '착하게 살면 윤리와 무관할 수 있다.'는 직관이 더 이상 통용되지 않는 것이다.

> 우리의 일상적인 언어 직관에 따르자면, '윤리'는 지극히 개인적인 사안에만 한정된다는 느낌이 있다.
>
> 이 직관은 표준국어대사전의 '윤리'에 대한 정의, "사람으로서 마땅히 행하거나 지켜야 할 도리"와도 일치한다.

우리가 사회적 수준에서 추구하는 여러 가치, 예를 들어 자유와 평등 사이에는 동시에 만족하기 어려운 긴장이 존재한다. 여러 가치를 최대한 동시에 실현하기 위해서는 이해당사자가 모두 완벽하게 만족하는 현실적으로 불가능한 방식이 아니라, 사회적 숙고를 통해 윤리적으로 합리적이라고 평가될 수 있는 방식으로 각각의 가치를 적절한 수준에서 절

충하여 만족하는 방식을 활용하게 된다. 당연히 AI 윤리의 여러 핵심 주제에 대해서도 마찬가지 방식으로 주요 사회적 결정이 내려질 수밖에 없다.

이렇게 이해된 AI 윤리의 관점에서 보자면 AI와 관련된 다양한 개인적, 사회적, 법적, 제도적 쟁점에 대해 단순한 선악 판단을 하려고 시도하기보다는 우리 사회에서 핵심적으로 존중되는 가치에는 어떤 것이 있으며 그 가치를 최대한 균형 있게 존중하는 방식으로 AI 개발과 활용을 하기 위해서는 어떤 점에 주의하고 어떤 제도적 장치를 마련해야 하는지를 통합적으로 탐색하려는 노력이 필요하다.[3,4,10]

이제부터는 이렇게 여러 가치를 통합적으로 고려하고 사회적으로 수용가능한 해결책을 찾아가는 과정이 정확히 무엇을 의미하는지를 보여주는 AI 윤리의 사례를 소개한다.

AI는 인간보다 더 공정할까?

이루다 사건으로 AI의 공정성이 사회적 관심사로 떠오르기 전까지만 해도 AI는 인간의 편견이나 사사로운 감정으로부터 자유롭기에 인간보다 훨씬 더 공정할 것이라는 생각이 지배적이었다. 유사한 사건에 대해 그때그때 기분에 따라 다른 형량을 부과할 수 있는 인간 판사 대신 객관적인 증거와 유사 사건의 판례만을 공정하게 조하여 판단할 수 있는 AI 판사에게 재판을 받고 싶다는 희망을 피력하는 사람도 있었다. 너무나 고려할 것이 많은 복잡한 의료 현장에서도 실수 없이 차분하게 정확한 진단이나 처방을 내리는 AI 의사를 사람 의사보다 더 신뢰한다는 의견이 언론에 보도되기도 했다. 하지만 이제 이루다 사건 이후로 사람들은 AI가 인간보다 더 공정할 수도 있지만 극단적인 방식으로 더 편견에 사로잡힐 수도 있다는 걸 알게 되었다.

그런데 정말 그럴까? 도대체 AI가 공정하거나 편견을 갖는다는 것은 정확히 무엇을 의미할까? AI가 공정해야 하는지에 대해 답하기 전에 이 문제부터 살펴보자.

현실의 AI vs. SF 영화 속 AI

AI와 관련된 윤리적 쟁점을 다룰 때 미리 분명하게 짚고 넘어가야 하는 점은 현실에 존재하는 (그리고 가까운 미래에 등장할) AI와 SF 영화에 등장하는, 인간과 구별되지 않는 수준의 감정 능력과 도덕적 판단 능력까지 발휘하는 가상의 AI 사이의 구별이다. 가까운 미래를 포함하여 당분간 우리가 경험할 AI는 인간의 특정한 능력을 '흉내'낼 목적으로 만들어진 특수지능이다. 이 사실은 중요한 함의를 갖는다.[2,8]

"

다른 사람과 협력해서 공동 작업을 수행하는 일 등을 인간 수준으로 해낼 수 있는 '일반지능'을 갖춘 AI는 아직 존재하지 않는다.

"

첫째, '이루다'와 같은 AI가 아무리 성차별적으로 간주될 수 있는 발언을 한다고 해도 AI는 상식적 의미에서, 성차별적 의도나 감정을 갖지 않는다. 실은 '성차별'을 포함하여 자신이 산출하는 문장들의 의미를 통상적인 의미에서 이해한다고 볼 수도 없다. 예를 들어 이루다가 산출하는 문장 기호를 우리가 읽고 이루다의 '마음 상태'를 유추할 뿐이지, 실제로 이루다가 의식적 마음을 갖고 있지는 않다.

둘째, 이루다를 비롯한 챗봇 AI가 지금보다 훨씬 더 발달해서 인간과 전혀 구별할 수 없는 수준의 대화를 나눌 수 있게 되더라도, AI가 평범한 인간이 하는 다른 일, 예를 들어 시각이나 음성을 통해 사람을 알아보거나 가게에 가서 물건을 사는 일까지 할 수는 없다. 물론 시각이나 음성을 통해 사람을 구별하거나 물건을 집거나 들어 올리는 일을 할 수 있는 AI 혹은 AI 로봇은 지금도 존재한다. 하지만 평범한 사람처럼 이 모든 일을 포함해 수많은 다른 일, 예를 들어 다른 사람과 협력해서 공동 작업을 수행하는 일 등을 인간 수준으로 해낼 수 있는 '일반지능General Intelligence'을 갖춘 AI는 아직 존재하지 않는다. 관련 연구조차 극히 초보 단계여서 가까운 시일 내에 우리 삶에서 일반 AI를 쉽게 볼 수 있을 가능성은 거의 없다.

AI '산출물'의 공정성

그러므로 이런 배경에서 AI의 공정성은 다음과 같이 이해해야 한다. 현재까지 (그리고 가까운 미래에 등장할) AI는 공정이란 단어의 의미도 알 수 없고 공정과 관련된 복잡한 의미론적, 사회적, 윤리적 관계를 이해할 수 있는 '의식적 마음'도 가질 수 없다. 그러므로 AI가 사람보다 더 혹은 덜 공정한가라는 질문은 이런 의식적 마음을 갖지 않은 복잡한 기계가 수많은 공학자들의 노력과 엄청난 양의 학습 데이터를 활용한 기계학습을 기반으로 산출하는 결과물이 사람이 보기에 동일한 일을 수행하는 사람이 산출한 결과물보다 더 혹은 덜 공정한가를 의미한다.

이렇게 정리하고 나면 처음 제기한 문제는 너무 쉽게 답할 수 있어 보인다. 결국 AI가 최대한 공정하게 결과 값을 내도록 잘 만들면 되지 않을까? 그런데 이 지점부터 문제가 복잡해진다. 본격적인 AI 윤리 논의가 시작되는 것이다. 우리는 AI의 결과 값이 공정한 것을 항상 원하는가? 조금만 생각해 봐도 우리가 AI를 활용하는 목적에 따라 그 답은 달라질 것 같다.

우선 공정이 무엇인지 생각해 보자. 표준국어대사전은 공정을 '공평하고 올바름'으로 정의한다. 핵심은 공정이란 개념은 평가적 혹은 규범적 개념이라는 점이다. 이 말이 무엇을 의미하는지 이해하기 위해 예를 들어보자. 여성의 '평균' 키는 남성의 '평균' 키보다 약간 작다. 이는 통계적 사실이고 이 사실을 말한다고 해서 성차별적이고 공정하지 않다고 말할 사람은 없다. 하지만 국내 100대 기업의 최고경영자 중에서 남성이 여성보다 압도적으로 많다는 점 역시 통계적 사실이지만 이 사실은 많은 사람들에 의해 성차별적이고 공정하지 않은 것으로 여겨진다. 차이가 뭘까? 이 두 사례를 비

교해보면, '공정함'이란 세상이 어떠하다는 사실적 주장과 관련된 것이 아니라 세상이 마땅히 어떠해야 한다는 규범적 주장과 관련됨을 알 수 있다. 대다수 사람들이 남녀 평균키가 같은 것이 윤리적으로 더 바람직하다고 보지 않는 반면, 남성과 여성이 비슷한 비율로 최고경영자가 되는 것이 윤리적으로 더 바람직하다고 보기 때문이다.

그런데 이렇게 정리해도 여전히 남는 문제가 있다. 최고경영자 중 남성에 비해 여성이 적은 이유는 실제로 '현재 기업 환경 조건'에서 남성이 여성보다 더 높은 성취를 보여주기 용이하기 때문일 수도 있다. 이 경우에는 남성과 여성이 최고경영자로서의 '잠재력'에 있어서는 평균적으로 완전히 동등하더라도, 기업 입장에서는 남성 최고경영자를 임용하는 것이 기업 '실제' 실적에 도움이 되기 때문에 남성 최고경영자를 선호할 수 있다. 이런 고려까지 하게 되면 결국 최고경영자 비율에서 남녀차이를 공정하지 않다고 지적하는 것은 '현재 기업 환경 조건'을 포함한 우리 사회 전체에 존재하는 여성에게 불리한 사회적 조건 전체에 대해 비판하는 것이 된다. 물론 이런 상황이 여성에게만 해당될 이유는 없다. 혹자는 현재 남성에게만 부여되는 병역의무가 남성에게 공정하지 않다고 주장하거나 남자에게 '남자다움'을 요구하는 우리 사회가 문화적 포용력을 결여하고 있다고 비판할 수 있다. 핵심은 '공정함'에 대한 규범적 판단이 명백한 성차별처럼 광범위한 지지를 얻을 수 있는 것에서부터 다소 논쟁적 사안까지 다양한 스펙트럼으로 존재한다는 사실이다.

"

결국 AI가 최대한 공정하게 결과 값을 내도록 잘 만들면 되지 않을까?

그런데 이 지점부터 문제가 복잡해진다. 본격적인 AI 윤리 논의가 시작되는 것이다.

"

그런데 여기서 잠깐 멈추어서 우리 사회의 공정하지 못한 '측면'을 파악하고 이에 대한 대응책을 마련하기 위해 AI를 활용하는 상황을 고려해 보자. 최근 사회정책 수립이나 사회문제 해결에 AI를 활용하면 좋을 것이라는 생각이 점점 인기를 얻고 있으니 충분히 가능한 상황이다. 이런 목적이라면 우리 사회에 어떤 불평등한 모습이 있는지를 가감 없이 그대로 드러내는 AI가 필요할 것이다. 이런 AI의 산출물이 보여주는 우리 사회의 불평등한 모습이 그에 대한 교정적 정책을 시행할 수 있는 정확한 출발점이 되어야 하기 때문이다.

이처럼 AI의 용도에 따라 AI의 공정성, 즉 AI 산출물의 공정성은 추구할만한 가치일 수도 있고 그렇지 않을 수도 있다. 특히 현재 사용되는 AI의 대부분이 현재까지 수집된 데이터를 기계학습하고 그 데이터 집합에서 발견되는 규칙성 혹은 패턴이 가까운 미래에도 성립할 것이라는 전제하에 미래를 예측한다. 이는 AI의 예측이 근본적인 수준에서 '보수적'일 수밖에 없음을 의미한다. 현재까지 행해져왔던 사회적 결정과 행동의 패턴이 미래에도 그대로 실현될 것이라는 가정을 깔고 있기 때문이다. 이런 목적으로 만들어지고 활용된 AI의 산출물이 공정하지 못하다고 비판하는 것은 그 자체로는 맞는 이야기지만 초점을 잃은 비판일 수 있다.

이제 질문을 좀 더 정교하게 가다듬어 보자. AI의 제작 목적에 따라 그 산출의 공정성을 요구하지 말아야 할 AI가 분명 존재한다. 그러므로 이런 종류의 AI를 제외하고 우리가 AI의 산출물의 공정함 자체를 요구해야 할 AI가 분명 있을 것이고 그에 대해 공정하기를 요구하는 것은 윤리적으로 바람직할 것이다. 이번에 문제가 된 이루다처럼 수많은 사람들과 미리 예측하기 어려운 방식으로 상호작용하는 사회적 대화형 AI의 경우에는 당연히 그런 공정성에 대한 요구가 강해질 수밖에 없다.

하지만 이 경우조차 사안은 여전히 더 복잡하다. 이루다와 같은 사회적 파급효과가 큰 AI에 공정함을 요구하는 것이 규범적으로 타당하다는 점에는 논란의 여지가 없지만 그때 요구되는 공정함이 어느 정도 수준이어야 하는지에 대해서는 사람들 사이의 직관이 쉽게 일치하지 않기 때문이다. 예를 들어 사람들에게 웃음을 선사할 목적으로 제작된 오락 프로그램에 지나치게 강력한 도덕적 잣대나 '정치적 올바름political correctness'을 요구하는 것은 오락 프로그램의 본질을 훼손하는 것이라는 비판이 제기되어 왔다. 우리는 챗봇 AI가 논란의 소지가 완벽하게 제거된, 물샐틈없이 '도덕적인 문장'만을 발화하기를 원하는가? 이 부분도 고민이 필요한 주제이다.

> 핵심은 AI가 단순히 공학자들이 만드는 기술이 아니라 우리 사회 전체의 윤리적 공감대를 반영해야 할 문화적 산물이라는 점을 인식하고 실천하는 것이다.

헌법이 보장하는 표현의 자유와의 충돌 문제도 있다. 물론 표현의 자유가 다른 모든 사회적 가치를 희생하면서까지 반드시 지켜야 할 절대적 가치는 아니다. 국제적으로 여러 국가들이 자국의 역사적, 문화적, 사회적 상황에 따라 특정 종류의 혐오표현에 대해 법적으로 처벌할 근거를 마련하고 있다. 그러므로 우리는 AI 설계 단계부터 표현의 자유와 다른 사회적 가치의 맞교환 문제를 고민해야 한다. 중요한 점은 이루다와 같은 공정함을 요구할 필요가 있는 AI의 기획 및 제작 단계에서 어느 정도의 '공정함'이 적절한 수준인지를 미리 고민하고 이를 알고리즘이나 데이터 수집 및 활용 과정에 반영하는 것이다.

정리해보자. 우리는 사회적 상호작용을 비롯하여 사람들의 행동이나 가치에 큰 영향을 끼치는 AI의 '산출물'이 공정할 것을 요구해야 한다. 그런데 어느 수준의 공정함을 요구해야 할지는 AI 제작 단계에서부터 충분한 학제적 논의를 통해 결정되어야 하고 이 결정 내용이 알고리즘 자체나 훈련 데이터의 수집 및 활용 과정에 반영되어야 한다. 핵심은 AI가 단순히 공학자들이 만드는 기술이 아니라 우리 사회 전체의 윤리적 공감대를 반영해야 할 문화적 산물이라는 점을 인식하고 실천하는 것이다.

이제 우리는 이 글 제목의 질문에 답할 준비가 되었다. AI 윤리란 무엇인가? AI 윤리는 (먼 미래에 등장할 일반 지능을 갖춘 AI를 배제하면) AI의 '산출물', 특히 인간의 지속적인 통제를 받지 않는 '자동화된 결정automated decisions'이 우리가 소중하게 여기는 기본 인권 등의 다양한 사회적 가치를 최대한 존중하는 방식으로 활용되기 위해서 어떤 점에 주목하고 어떤 방식의 제도적 대응을 수행해야 하는지에 대한 논의이다. 그리고 이 논의와 그로부터 파생되는 제도적 실천은 AI 개발과 활용의 전주기entire lifecycle에 적용되어야 하고 그 논의가 영향을 미치는 집단의 윤리적 공감대와 문화를 적극적으로 고려해야 한다.

참고문헌

1. 이상욱, 조은희 엮음 2011,『과학 윤리 특강 – 과학자를 위한 윤리 가이드』, 서울: 사이언스북스.
2. 이중원 외 2018,『인공지능의 존재론』, 서울: 한울.
3. 이중원 외 2019,『인공지능의 윤리학』, 서울: 한울.
4. 한국인공지능법학회 2019,『인공지능과 법』, 서울: 박영사.
5. Fry, Hannah 2019, *Hello World: How to Be Human in the Age of the Machine*, London: Transworld Publishers Ltd.
6. Gans, Joshua, Goldfarb, Avi and Agrawal, Ajay 2018, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Cambridge, MA: Harvard Business School Press.
7. IEEE 2019, *Ethically Aligned Design*, 1st Edition. (https://ethicsinaction.ieee.org/#series 참조)
8. Kaplan, Jerry 2016, *Artificial Intelligence:What Everyone Needs to Know*, Oxford:Oxford University Press.
9. Mitchell, Melanie 2020, *Artificial Intelligence: A Guide for Thinking Human,* New York: Picador.
10. Sharre, Paul 2019, *Army of None: Autonomous Weapons and the Future of War*, New York: W.W. Norton & Co.
11. Susskind, R. and Susskind, D. 2017, *The Future of Professions: How Technology Will Transform the Work of Human Experts*, Oxford:Oxford University Press.
12. UNESCO 2019, *Preliminary Study on the Ethics of Artificial Intelligence*
13. UNESCO 2020, *First Draft of the Recommendation on the Ethics of Artificial Intelligence*