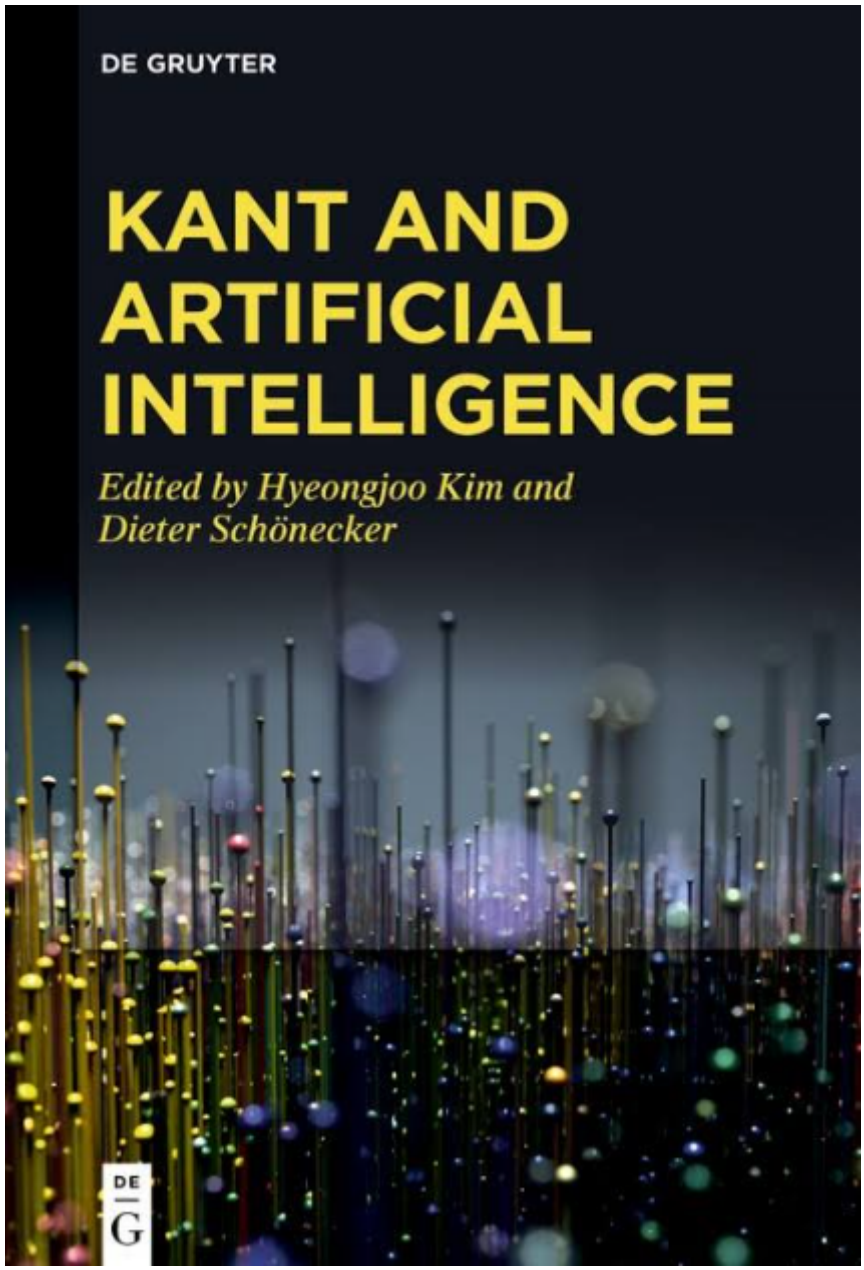
칸트와 인공지능

김형주 중앙대·철학 | 승인 2023.06.17 19:17

■ **Hyeongjoo Kim & Dieter Schönecker (eds.), 2022. *Kant and Artificial Intelligence*. De Gruyter**



진리가 너희를 자유케 하리라

해석은 자유다. 요한복음 8장 31절 '진리가 너희를 자유케 하리라'는 다방면의 자유로운 해석에 가장 많이 노출된 구절 중 하나일 것이다. 신학자들은 진리를 예수로, 자유를 죄로부터의 자유로 번역한다. 교회는 이 구절에 의지해 예수가 구원자임을 믿으면 죄로부터 자유를 얻는다고 가르친다. 이 구절은 미국의 중앙정보국 CIA의 헤드쿼터의 한 벽면에도 새겨져 있다. 이는 CIA의 비공식 모토인데, CIA가 여기에 부여한 의미는 정보가 자유로운 사회를 보장한다는 것인 듯하다. 세계의 각 대학들도 이 구절을 앞다퉈 사용한다. 대표적으로 우리나라의 연세대, 독일의 프라이부르크 대학이 이 구절을 교훈으로 사용하는데, 이때 이 구절에는 '학문과 사상의 자유'라는 상징이 덧입혀진다.

한편, 나와 같은 칸트 연구자는 이 구절에서 '자연의 법칙'과 '자유의 법칙'의 조화를 읽는다(IV387, V175). '진리'에서는 우리의 인식에 조응하는 객관 세계의 법칙을, '자유'에서는 자연의 법칙으로부터 자유로운 예지계의 법칙, 즉 자율도덕의 법칙을 읽어낸다. '진리가 너희를 자유케 한다'는 말은 적어도 우리 칸트 연구자에게는 자연의 법칙과 자유의 법칙의 통일로 읽힌다. 선험 철학으로 명명되는 칸트의 인식론은 자연과 세계를 구조화하는 인간 이성의 체계를 밝힌다. 지구상에서 유일하게 자유로운 존재자인 인간이 스스로 만들어 스스로에게 부여하는 법칙인 도덕의 법칙은 칸트의 윤리학이 발견한 성과이다. 칸트의 철학은 자연과 자유의 철학이다.

자연과 자유의 철학의 과제는 종국에는 인간 존재의 근거 물음으로 귀결되는 세 가지 물음, '나는 무엇을 알 수 있는가?', '나는 무엇을 해야 하는가?', '나는 무엇을 희망해도 좋은가?'로 집약적으로 제시된다(A805/B833). 내가 알 수 있는 것은 내 눈으로 보아 나의 기억에 남겨진 것, 즉 현상이지 사물 자체가 아니다. 내가 해야 하는 것은 자유의 법칙에 자율적으로 복속하여 인간의 존엄성을 지키는 것이다. 내가 희망해도 될 것은 "공기의 저항에서 벗어난 비둘기의 날갯짓"이 아닌, 도덕적 삶을 다 살아내고 난 후 얻게 될 행복에 대한 기대감이다. 우리는 이 자연과 자유의 시각으로 요한복음의 경구를 읽어내었듯, '인공지능'을 바라보았다. 그 결과가 'Kant and AI'이다.

칸트가 본 인공지능(Artificial Intelligence from Kantian Perspective)

이러한 배경에서 자연의 철학, 즉 이론철학이 첫 장에 놓인다. Deep Mind의 인공지능 공학자이자 이제는 칸트 연구자라고도 할 수 있는 리차드 에반스가 개발한 칸트 머신 Apperception Engine이 화두를 던진다. 그는 순수이성비판에서 칸트가 정치하게 기술한 인간 인식의 원칙을 비지도 학습 인공지능 아키텍처로 구현하였다. 감성과 지성의 협주를 통해 산출된 경험을 그는 'sense making'으로 표현하면서 직관을 종합하는 감성, 감성이 포섭한 자료들을 판단하는 지성의 범주의 원리를 프로그래밍하였다. 그가 주목한 것은 칸트 인식 모델의 가장 근원적인 개념, 칸트가 기술한 인간 인식 능력 중 지성에서 발원하였지만 감성의 최초 활동에도 작동을 하는 종합 활동이다.

이어 영국 Keele 대학의 Sorin Baiasu 교수는 "the Challenge of (Self-)Consciousness: Kant, Artificial Intelligence and Sense-Making"에서 "이 칸트 머신의 성과가 인간과 비견할만하다"라고 말하면서(the performance of the Kantian machine here is comparable to that of human beings, 107쪽) 원본적 지향성을 가진 인식 행위자로 간주할 수 있다고 평가한다. 외부 세계에 대한 표상을 종합하는 능력이 인간 인식종합의 원리, 그렇다고 칸트의 인식 모델이 빠짐없이 구현된 것은 아니라고 말한다. 즉, 규칙을 통한 종합은 필요조건이지 아직 충분조건은 아니다(combining through rules is necessary for my experience, but it is not sufficient, 126쪽). 필요조건과 충분조건 사이의 갭을 메우기 위해 칸트는 여전히 자기의식을 요청한다. 에반스도 이를 부정하지 않는다. 그는 비록 통각 엔진이 통각의 종합적 통일, 즉 인풋 데이터들의 종합을 목적으로 하였지만, 결코 통각의 분석적 통일, 즉 그 데이터들이 바로 내 인식의 것이라는 의식의 구현을 시도하지 않았다고 말한다(99쪽).

한편, 독일 보훔 대학의 토비아스 쉴리흐트(Tobias Schlicht) 교수는 칸트를 현대 인지과학의 토대를 마련한 "지성적 대부(intellectual godfather, 3쪽)"로 규정한 앤드류 브룩(Andrew Brook)의 언급을 시작으로 칸트의 공적을 심층신경망 이론(Deep Neural Networks)에까지 연결시킨다. 그는 기능주의(functionalism), 구성주의(enactivism)에 미친 칸트의 영향을 소개하고 기계학습 모델 개발에 큰 영향을 준 예측 처리(predictive processing)와 칸트의 인식론의 연관성을 칸트의 인식론을 특징짓는 표현인 '코페르니쿠스적 전회'를 통하여 논증한다. 나아가 딥러닝 기술의 가능성과 한계를 고찰하고 인공지능 개발의 최종 목표라고 할 수 있는 범용 인공지능 실현의 가능성을 오랜 철학 논쟁인 경험론과 합리론의 대결을 기반으로 낙관주의(optimist)와 회의주의(pessimist)를 두루 살펴본다. 이때 인과율과 같은 지성의 선험적 범주의 인정 여부가 중요한 주제로 등장한다. 논의는 다시 통각 엔진으로 돌아오는데, '지성의 자기 의식적 활동'(B153)의 부재가 이론적 공백으로 제기된다.

이들이 최신 인공지능 기술에 주목했다면 김형주는 인공지능의 본래적 개념에 주목한다. 김형주는 인공지능 개념의 창시자인 존 맥카시(John MacCarthy)의 문건들을 분석하여 인공지능의 본래적 의미가 경험적 물리적 세계의 문제 해결을 위하여 인간 지능의 계산 능력을 모방한 지능이라 규정하는 한편, 칸트가 자신의 세계관을 특징짓는 개념인 선험적 관념론(transcendental idealism)과 이를 둘러싼 개념들(A369)을 범주화하여 인공지능 세계관을 선험적 실재론에 위치시킨다. 김형주도 여전히 에반스, 쉴리히트, 바이어수와 마찬가지로 표상들의 종합능력을 인공지능과 인공지능의 공통점으로, 자기의식의 소유 여부를 양자의 차이점으로 지적한다.

사진 출처: asianews network

2장은 자유의 법칙을 탐구하는 실천철학적 논의로 구성되어 있다. 주지하듯, 칸트는 윤리학을 두 가지로, 도덕형이상학과 실천적 인간학으로 나눈다. 현대적 관점에서 보자면 전자는 윤리학에. 후자는 응용윤리에 상응할 것이다. 한편 논의의 폭을 넓히면 실천철학에는 사회, 정치 철학도 포함된다. 독일 지겐 대학교의 디터 쇠네커 교수는 칸트의 도덕 감정 개념을 근거로 하여 인공지능의 존재론적, 윤리적 지위 문제를 다룬다. 그는 도덕 감정, '나는 생각한다', '나는 느낀다'에 대한 칸트적 설명에 대한 정치한 분석을 행한다. 이를 바탕으로 칸트적 관점을 취하면 인공지능은 감정의 주체가 느끼는 주체가 될 수 없다는 주장을 펼치는데, 결론부에서 수영하는 잠수함(Swimming Submarines)의 비유는 존 설과 앨런 튜링의 논쟁을 함축하는 중국어방 논쟁의 도덕철학적 버전이다. 수영한다는 것을 물속에서 움직인다는 것으로 정의한다면, 컴퓨터도 생각한다고 말할 수 있다. 그러나 이 유비는 잘못되었다. 어떤 단어를 정의하려면 그것의 본질에 대한 인식이 전제되어야한다. 어떻게, 무엇을 통해 수영이라는 것을 행하는지 묻는 것은 수영의 개념과는 상관이 없다. 인공지능이 느낀다는 말은 비행기가 공간을 이동하기 때문에 난다고 말하는 것과 같다. 쇠네커 교수는 생각, 감정 욕구의 담지체(Substratum)인 자아(I)가 없는 인공지능은 윤리적 주체일 수 없다는 강경한 입장을 제기한다.

이러한 입장은 스탠포드와 쾰른 대학의 Lisa Benossi와 Sven Bernecker의 논의로 이어진다. 하지만 이들의 논의는 조금 더 현실적인 문제로 향한다. "A Kantian Perspective on Robot Ethics"에서 던지는 질문은 '완전한 도덕적 행위자로서의 인공지능 로봇이 개념적으로 가능한가'에서 한 걸음 더 나아가 "도덕적 행위자인 로봇을 개발하는 것을 규제하는 것은 도덕적으로 정당한가"로 향한다. 칸트가 규정하는 완전한 도덕적 행위자는 스스로 보편적 행위의 원리에 부합하도록 자신의 행위의 원리를 수립하고 이에 스스로 복종하는 존재다. 그렇기에 그는 다른 모든 이성적 존재자들과 동일한 권리와 의무를 갖는다(149쪽). 따라서 "우리는 그런 존재를 단지 수단으로서가 아니라 동시에 목적으로도 대우해야 한다."(IV433). 만약 로봇이 그러하다면 우리는 로봇에게 우리가 싫어하는 일이나 위험한 일을 시킬 수 없게 된다. 로봇이 도덕적이 될수록 유용성은 떨어진다.

칸트의 윤리학을 아리스토텔레스의 잠재성 논변을 빌려 폭넓게 해석하면, 어린 아이, 심지어 동물도 도덕적 고려의 대상이 된다(163-164쪽). 이러한 존재들의 도덕성 발달을 제한하는 것은 도덕적으로 부당하다. 하지만 로봇은 자기 스스로를 예지계의 성원으로 복속시킬 도덕 법칙을 산출할 수 있는 존재가 될 수 없기에, 도덕적 행위자성을 갖는 로봇의 개발을 제한하는 것은 정당하다. 한편, 자신뿐 아니라 다른 모든 사람의 인간성을 수단으로뿐 아니라 목적으로 사용하라(IV433)는 정언명령의 인간성 정식을 취한다면, 개발의 범위를 추론할 수 있다. 그 바로미터는 아시모프의 3원칙이 말하듯 바로 인간의 존엄성이다.

사진 출처: PHILOS-SOPHIA INITIATIVE

이상의 고찰들이 인공지능의 도덕적 지위 문제를 다루는 이른바 좁은 의미의 윤리학적 논의였다면 자율주행차의 윤리적 이슈를 다루는 엘케 엘리자베스 슈미트(Elke Elisabeth Schmidt)와 에바 토마스 라이트(Ava Thomas Wright)의 논의는 응용윤리학적 논의라고 할 수 있다. 트롤리 딜레마를 다룸에 있어 이 문제를 처음 제기한 필리파 풋(Philippa Foot)은 이 문제를 해결하는 두 축으로 소극적 의무(사람을 죽이지 말라)와 적극적 의무(다른 사람을 도와라 혹은 살려라)를 이야기한다. 슈미트와 라이트는 이를 위해 칸트의 개념을 끌어온다. 슈미트는 소극적 의무와 적극적 의무를 각각 좁은 의미의 의무(narrow duty)와 넓은 의무(wide duty)로, 라이트는 법적 의무(duty of right)와 윤리적 의무(ethical duty)로 연결 짓는다. 풋이 제기한 트롤리 딜리마의 오리지널 버전은 다음과 같다. 곧 갈림길을 만나는 외길을 달리고 있는 트롤리가 있다. 외길의 앞에는 5명이 있고 갈림길의 끝에는 1명이 있다. 스위치를 당겨 선로를 바꾸면 1명이 죽고 그렇지 않으면 5명이 죽는다. 풋은 '소극적 의무와 적극적 의무가 충돌할 때는 소극적 의무가 우선한다', '소극적 의무가 충돌할 때는 희생자를 최소화해야 한다'(191쪽)는 두 가지 원칙을 제시하면서 선로를 바꿔 한 명을 희생시키는 것이 윤리적이라는 결론을 내린다.

슈미트는 칸트주의자가 완전자율주행 단계의 자율주행차의 프로그래머라면 어떤 원칙을 세울지에 대해 논의하기 위해서는 차를 직접 조정하지는 않지만 스위치를 조정할 수 있는 주체가 행인인 행인의 예가 논의의 대상이 되어야 한다고 주장한다. 칸트의 입장을 취하면 풋과는 다른 결론이 도출된다. 행인은 사람을 죽이지 말라는 좁은 의미의 의무에 복종해야 한다. 슈미트는 이러한 결론에 대한 반대 주장을 차례로 논박한다. 라이트도 슈미트와 마찬가지로 좁은 의미의 의무, 즉 법적 의무의 우선성에 주목하는 한편, 자율주행차 논쟁을 특수한 상황에서 마주하는 도덕적 딜레마에 대한 규범 윤리적 해결책 모색으로부터 "옳은 행위(right)는 보편적 법칙 아래 모든 사람의 자유와 일치할 수 있는 행위"(VI230)라는 칸트의 원칙을 토대로 법의 공적 권위와 거버넌스, 공적 체계 차원의 논의로 확장시

킨다. 이와 관련하여 최근 OECD, EU등 국제기구에서 발간하는 인공지능 윤리 정책보고서의 논조도 윤리 가이드라인 제시에서 구속력과 강제력을 가진 법적 차원의 구체적 지침으로 변하고 있다는 사실도 생각해 볼만하다.

라이트가 제시한 사회철학적 관점은 튀빙겐 대학 클라우스 디르크스마이어(Claus Dierksmeier) 교수의 논의로 이어진다. 앞에서 본 것처럼 칸트의 관점을 취하면 인공지능 기술 발전에 대해 다소 보수적인 입장을, 기술 그 자체에 대해서도 비판적인 입장을 취하게 된다. 이와는 달리 디르크스마이어는 칸트의 유기체 철학의 관점에서 긱 경제(gig economy)를 배경으로 새롭게 등장한 인공지능 기반 직업 매칭 시스템을 긍정적으로 해석한다. 그는 스타트 업 기업이 개발한 어플리케이션인 'TiiQu'와 'YourCompany'를 인간의 자율성을 향상시킬 수 있는 새로운 플랫폼이라고 소개한다. 그는 칸트의 '목적의 나라' 개념에 주목한다. 주지하듯 '목적의 나라'는 "공동의 객관적인 법칙에 의한 이성적인 존재들의 체계적인 결합"(IV433)으로 기계적인 인과성(mechanic causality)만이 아닌 목적론적 인과성(teleological causality)의 영향을 받기에 이를 구성하는 존재자들은 일부(part; Teil)가 아니라 구성원(limb; Glied)이다(241, 242쪽).

그는 목적 그 자체인 개체들이 상호 영향을 받으며 자연을 구성하고 자신의 내적인 목적을 실현하는 과정을 그린 유기체 철학과 사회 조직을 유비한다. 이러한 유비로 비추어 볼 때 칸트적 의미의 사회 조직(social organization)은 그 구성원들(limbs)이 서로 영향을 주고받으며 도덕적 주체로서의 자기의 목적, 즉 자신의 자율성을 증진시키는 조직이라고 할 수 있다. '직업 매칭 플랫폼 TiiQu'와 'YourCompany'의 고용자, 지원자의 관계는 수평적으로 상호 영향을 주고받으며 개인의 자율성을 증진시키는 관계이다. 특별히 이 플랫폼을 구성하고 있는 알고리즘은 신뢰할 수 있는 고용주와 정직한 직원 모두에게 이익이 되도록 채용 프로세스와 매칭 시스템을 익명화하고 표준화하며, 고정관념, 연고주의, 편애(biases, stereotypes, nepotism, favoritism)와 같은 것이 고용에 영향을 미치는 것을 차단한다(247쪽). 이러한 의미에서 이 새로운 노동 플랫폼은 칸트적 의미의 자율적이고 유기적인 사회 조직이다. 한편, 그는 이 플랫폼이 준 공적 권력(public power)을 갖고 있으므로 공적인 감시체계를 갖추는 것을 간과하지 말아야한다고 칸트의 입을 빌려 주장한다(250쪽).

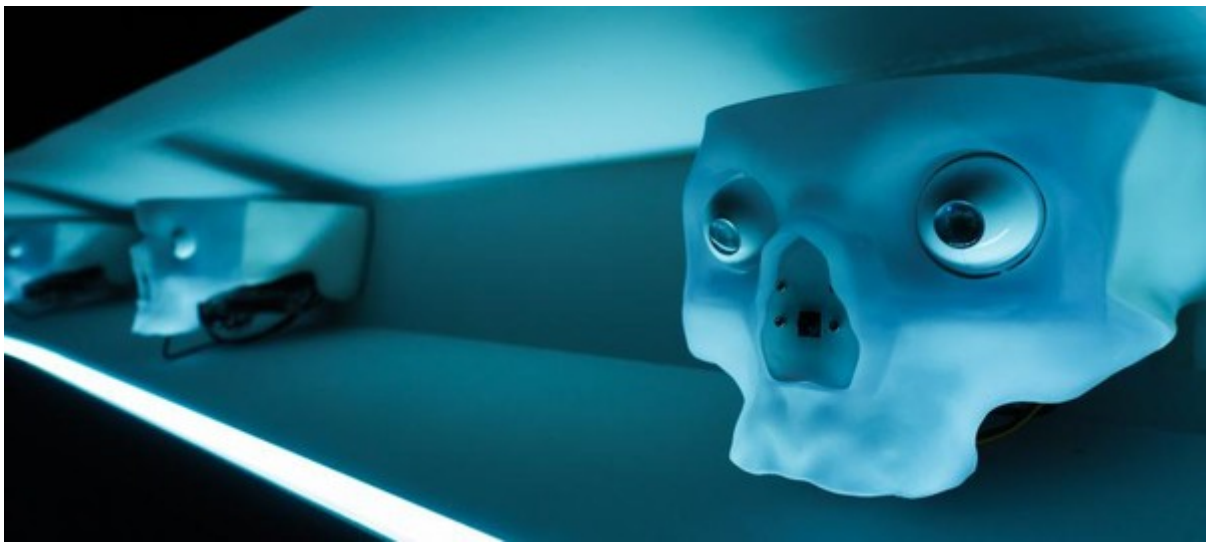


사진 출처: Philosophy Talk 블로그

진, 선, 미

1787년, 칸트는 자유의 철학『실천이성비판』을 탈고하고 그의 지적 동료 라인홀트(Reinhold)에게 다음과 같이 말한다. "나는 지금 종전의 것들과는 다른 새로운 종류의 선험적 원리들이 발견된 것을 계기로 취미 판단에 몰두하고 있다. 무릇 마음의 능력은 셋이 있으니, 인식능력, 쾌, 불쾌의 능력, 욕구능력이 그것이다. 나는 첫째 것을 위해 순수 이성비판에서, 셋째 것을 위해서 실천이성비판에서 선험적 원리들을 찾아냈다. 나는 둘째 것을 위해서도 그것들을 찾았다"며 판단력비판의 집필 의사를 밝힌다. 우리는 자유의 철학을 '진(眞)'으로, 자유의 철학을 '선'(善)으로, 쾌와 불쾌의 철학을 '미(美)'의 철학으로 분류한다.

MIT의 라리사 베르거는 칸트 철학체계의 마지막, 미학의 관점에서 인공지능을 바라본다. 그녀는 미국의 현대 심리철학자 네이글(Thomas Nagel)의 일인칭 관점의 경험성(the first-person experience)을 토대로 한 주관성 개념을 칸트의 아름다움에 대한 쾌와 연결시킨다. 인간의 미감이 개인 내면의 현상적 삶에 근거를 둔 반면, 컴퓨터는 객관적 관점, 물리학적 법칙에 구속되어 있기 때문에 그것은 미감의 주체가 될 수 없다는 것이 그녀의 주장의 요지이다. 그녀는 그 근거를 쇄너커 교수와 마찬가지로 느낄 수 있는 능력의 부재에서 찾는다. 쇄네커 교수가 논증하였듯, 도덕 감정의 부재가 인공지능이 윤리적 주체가 될 수 없다는 것의 근거였다면, 쾌(Pleasure)를 느낄 수 없음이 미적 주체가 될 수 없다는 것의 근거이다.

이렇게 우리는 칸트의 렌즈를 '진, 선, 미'의 안경에 삽입하여 인공지능을 바라보았다. 바다 끝 수평선이 바다의 끝을 가리킴과 동시에 창공의 시작을 알리듯이, 인공지능을 바라보는 칸트의 시선은 기술 발전의 한계점을 시사함과 동시에 더 좋은 우리 삶을 위한 규제적 이념을 제공한다. 거친 요약으로 인해 필자의 의도가 훼손된 것은 아닌지 걱정된다. 그래도 글을 시작하면서 의지한 한 가지 사실에서 위안을 찾는다. 해석은 자유이다.

김형주 중앙대·철학

중앙대학교 인공지능인문학단 HK 교수. 칸트를 주제로 독일 지겐대학교에서 철학 박사 학위를 취득했다. 2016년 알파고로 시끄럽던 당시 우리 사회의 인공지능 담론장에「인간 지능과 인공 지능 개념에 대한 철학적 분석 시도」라는 논문으로 처음 말을 보탰다. 독일 주정부 프로젝트 Bewusste KI(인공지능과 의식)에도 참여했다. 독일, 미국, 영국의 칸트 연구자들과 함께『Kant and AI』를 집필했고, 대한민국 AI윤리원칙 수립에 참여했다. 2023년 지겐대학교 Bollenbeck Professor Fellowship에 선발되어 방문교수직을 수행할 예정이다.

김형주 중앙대·철학