https://doi.org/10.33156/philos.2022.65..008

인공이성비판의 가능성 물음*

김 형 주**

주제분류 응용 윤리학, 독일철학, 비판이론

주 요 어 인공지능, 인공지능 윤리, 인공이성비판, 칸트, 딜타이, 호르크하이머

요 약 문

이 연구는 이성이 감행하는 인공지능 시대에 대한 비판을 인공이성비판으로 규정하고 그것의 학문적 가능성을 타진해 보는 것을 목표로 한다. 이를 위해 칸트, 딜타이, 호르크하이머의 이성비판 기획을 차례로 검토하면서 이성비판 기획의 계보를 개략적으로 구성해 본다. 이를 통해 칸트의 이성비판 개념이 각기 다른 시대에 어떻게 자기 자신을 드러내고 있는지 밝힐 것이다. 그 과정에서 이성비판의 의미에는 이성의 자기비판으로서 이성을 구성하는 요소들에 대한 분석, 이성의 타자 비판으로서 당대의 학문 비판, 나아가 학문과 교통하고 있는 시대정신에 대한 비판이 속해 있다는 사실, 이성비판 계보의 근본정신은 '이성의 자율성 복권', '체계에 대한 추구'라는 사실도 논증할 것이다. 이러한 결과들을 현재 유행하고 있는 인공지능 인문학에 적용한다. 구체적으로 말해 비판의 첫 번째 역할에 따라 인공지능 인문학을 인공지능 철학, 포스트휴머니즘, 디지털 인문학, 인공지능 윤리로 범주화한다. 그리고 '체계에 대한 추구', '자율적 이성의 복권'이라는 '이성비판'의 시각에서 이렇게 범주화된 인공지능 인문학을 비판해 보면, 이들의 공통된 특성으로 경험주의적 양화주의가 도출된다고 논증한다. 마지막으로 이러한 입장이 인문학을 포함한 모든 학문 단위에 천편일률적으로 적용되는 것은 바람직하지 않다고 주장한다. 그 대안으로 재귀적 인간중심주의를 제시한다. 재귀적 인간중심주의는 인간의 실존적 존재 긍정과 인식론적 한계를 수평적으로 받아들인 이성의 솔직한 자기

* 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017S1A6A3A01078538)을 밝힙니다.

** 중앙대학교 인문콘텐츠연구소 HK교수

고백이라 간주하여 이를 인공지능 윤리가 지향해야 할 새로운 기조라고 평가한다.

1. 지금은 인공지능 시대

"우리는 기술을 장악하길 원한다. 이 기술 지배의 욕구는 기술이 인간의 통제력에서 벗어날 무렵에 다다를수록 더 급박해진다."[1] 기술철학자 하이데거가 1950년대 라인 강에 세워진 수력발전소를 바라보며 뱉은 이 고백은 강인공지능, 특이점, 초지능과 같은 말들이 유행하는 오늘날에도 여전히 유효하다.

섣불리 인정하길 꺼려 했던 사실, 지금은 '인공지능 시대'다. 선언적 표현에 불과하였던 '인공지능 시대'의 청사진은 이제는 현실이 되었다. 구글(Google), 페이스북(Facebook)과 같은 국제 기업은 세계 데이터 전쟁을 주도하고 있다. UN과 UNESCO와 같은 국제기구는 한 달이 멀다하고 인공지능 관련 정책보고서를 쏟아내고 있으며, 일본은 국가 주도로 디지털 청 설립을 추진하는 한편, 중국은 한발 앞서 인공지능 초등교육 전면 실시를 천명하였다. 우리나라도 이에 뒤질세라 인공지능을 미래 먹거리 창출의 첨병으로 선언하고 디지털 뉴딜 정책을 추진하고 있다. 과학기술정통부 산하 여러 연구소들은 인공지능과 관련한 정책보고서를 끊임없이 생산하고 있다. 하버드 대학의 버크만 클라인 연구소(Berkman Klein Center)는 2020년 발간한 보고서 "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI"[2]에서, OECD와 같은 국제기구, Google, IBM과 같은

1) Heidegger, M. (2000). "Die Frage nach der Technik", *Gesamtausgabe Bd. 7,* Vittorio Klostermann, Frankfurt am Main, 8쪽. 그는 이러한 인식아래 당시를 원자력 시대로 규정한다. Heidegger, M. (1997). "Der Satz vom Grund", *Gesamtausgabe Bd. 10,* Vittorio Klostermann, Frankfurt am Main, 45쪽.

2) Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI.* Berkman Klein Center Research Publication, (2020-1).

국제 기업, IEEE와 같은 학술단체가 2016년부터 2020년까지 AI 윤리 원칙과 관련하여 발간한 문건이 총 36개에 달한다고 밝히고 있다. 더욱이 이 문건의 발간 횟수가 2016년에는 2건이었던 것이, 2017년에는 4건으로, 2019년에는 14건으로 가파르게 늘었다는 사실은 그 유행의 정도가 점차 심해지고 있다는 사실을 증명한다. 인공지능 시대는 시작되었다.

항상 그래왔듯, 인공지능을 위시한 기술 문명에 대한 철학적 거리두기는 테크노포비아(techno-phobia)에 의한 냉소적 부정이 아닌 방향설정의 학문으로서의 소임을 수행하기 위한 본성적 태도라 할 수 있다. 이 글도 이러한 시도 중 하나이다. 이 글은 크게 세 부분으로 나누어진다. 첫째, 인공이성비판이 표방해야할 이념 구성에 대해 논구한다. 인공이성비판을 풀어 설명하면 이성이 감행하는 인공지능 시대에 대한 비판이다. 이를 위해 시대와 학문에 대한 기초놓기 작업이라 할 수 있는 '이성비판' 작업의 선례를 살핀다. 이는 '이성비판'의 계보학 구성 시도라고 할 수 있다. 그 대상은 18세기 독일 관념론의 시초 칸트의 순수이성비판, 19세기 정신과학의 정초 딜타이의 역사이성비판, 그리고 20세기 기술-자본주의 비판의 정초 호르크하이머의 도구적 이성비판으로 삼는다. 또한 시대정신 이해의 통로로서의 학문 체계에 대한 분석, 이들을 한데 아우르는 공통의 이념 제시를 계보학적 분석의 결과로 도출한다. 둘째, 글의 시작에서 언급했듯, 현시대를 인공지능 시대로 규정한 것을 전제로 인공지능과 관련된 담론을 '인공지능 인문학'으로 이름하여 그 요소들을 분석하여 이를 범주화한다. 마지막으로 인공이성비판의 역할을 약술한다. 그 역할은 인공지능 시대의 인문 담론, 즉 '인공지능 인문학'을 관통하고 있는 공통적 세계관을 분석적으로 드러내고, 이 글이 견지하고 있는 '이성비판'의 관점에서 이에 대한 비판적 가능성을 제기하는 것이다. 나는 그 미래적 가능성을 역설적이지만 학문에 대한 토대물음과 인간중심주의에 천착했던 근대성에서 찾는다.

2. ‘이성비판’의 철학과 ‘인공이성비판’

『순수이성비판』이 지성사에 변곡점을 형성한 이후, ‘이성비판’이라는 이름으로 시행되고 있는 연구의 역사는 지속되고 있다. 이성의 자기 복권 기에 칸트(Kant)의 ‘순수이성비판’이 이성비판을 개시하였다면, 자연과학의 패권확장에 따른 인문학의 실증화 시기에는 딜타이(Dilthey)의 ‘역사이성비판’이, 인간성 소외를 초래한 자본주의와 물질문명 팽창기에는 호르크하이머(Horkheimer)의 ‘도구적 이성비판’이 그 역할을 감당하였다고 할 수 있다.[3] 이 장은 세기를 거듭하며 이성비판의 이름으로 진행된 이 이론들의 상호관계에 주목한다.

이성의 자기비판과 시대비판은 근대적 이성이 시작하였다는 것은 주지의 사실이다. 『순수이성비판』을 비롯하여 세 비판서 시리즈를 저술한 칸트는 ‘이성비판’의 의미를 일차적으로 이성의 자기 능력에 대한 정확한 이해를 통한 한계 인식으로 규정한다.[4] 정확한 이해를 위해서는 분석이 선행하여야 하고 한계 인식 다음에는 이성의 부당한 사용에 대한 비판이

3) 최근에는 회슬레(Hösle)의 『이해적 이성비판』(Kritik der verstehenden Vernunft), 포스트휴머니즘의 대명사인 슬로터다이크(Sloterdijk)의 칸트적 계몽주의 비판인 『냉소적 이성비판』(Kritik der zynischen Vernunft)도 ‘이성비판’의 이름으로 전개되었다. 엄밀한 학문적 태도를 견지한다면, 1700년대 후반 칸트로부터 1930년대 비판이론가에를 아우르는 이른바 비판 계보학 구성의 가능성에 대해 회의적인 입장을 취하는 것이 안전할 것이다. 예컨대 비판이론에서의 비판은 사회 구조에 내재되어 있는 모순 대한 비판을 의미하는데 비해, 코젤렉(Kosellek)의 칸트에 대한 평가를 빌려 말하자면, 칸트에게 있어서 비판은 대상에 대한 주관적 판단 양식이기에 양자를 한데 엮어 말하기에는 어려움이 따른다고 말할 수 있다.(정대성, 2012) 그러나 딜타이와 호르크하이머가 칸트의 비판 개념을 계승한 흔적을 자신의 저서 여러 곳에 남겨두었다는 사실이 이러한 시도에 대한 원천적인 불가능성은 해제시켜준다. 이 절에서 나는 이러한 작은 편린들에 의지하여 이성비판의 계보학의 윤곽을 구성하고자 한다.

4) Kant, I.(1900 ff). Kritik der reinen Vernunft, Berlin: Walter de Gruyter, B XIX ff.

따른다. 이러한 의미에서 '이성비판'의 첫째 요건은 자기인식이다. 이를 위해 이성은 체계를 구성하여 자신을 바라본다. 그 결과 감성, 지성, 상상력, 욕구 등과 같이 자신을 구성하는 부분들을 건축술적(architetonisch)으로 분석하고 구조화한다. 그렇기 때문에 '이성비판'은 '이성의 자기비판'의 의미를 갖는다. 여기서 또 다시, 비판은 첫째 비판의 예비학으로서 분석학을 의미하고, 둘째, 그것의 완성은 자기 능력의 장점과 한계에 대한 인식이라 할 수 있다. 한편, '이성비판'은 이성이 행하는 학문 비판을 포함한 시대비판을 의미한다. 이것이 '이성비판'이 갖는 두 번째 의미이다. 칸트 당시의 철학은 칸트 스스로가 평가하길 "끊임없는 전쟁터"였다. 그에 따르면 세계지(Weltweisheit)로서의 철학은 설득력있는 세계관을 제시해야 하는데, 당시의 철학은 경험주의에 쓴 뿌리를 둔 회의주의와 이성에 대한 맹신과 남용에 의거한 합리주의적 독단론의 경주로 인해 종교혁명과 과학혁명에 따른 세계상의 변화에 적실하고 건설적인 답을 제시하지 못하고 있었다. 변동기의 세계상에 칸트는 자연과학이 보여준 학문적 엄밀성을 수용하여 선험철학을 주창하였고 이를 폿대 삼아 18세기의 철학은 물론 삶과 정치와 종교를 비판하였다. 이 비판 정신은 '자율성 논증을 통한 인간의 존엄성 확보', '원칙과 체계의 본질 밝힘을 통한 삶과 행위의 원리 규명'과 같은 근대의 세계관의 중심이념을 형성하기에 이른다.

순수이성비판으로부터 약 한 세기가 지난 후, "칸트의 비판의 길을 따르는 것"[5]을 당시 철학의 과제라고 못을 박았던 딜타이는 "정신과학에 대한 모든 숙고의 근본과제"[6]로서 "역사이성비판"이라는 개념을 주창한다. "역사이성비판" 개념이 주제적으로 다루어지는 곳은 1910년 그의 서거 직전 저술이자 대표작인 『정신과학의 역사적 세계의 건립』(Der Aufbau

5) Dilthey, W. (1957 – 2006). Gesammelte Schriften (앞으로는 줄여서 GS로 표시), Stuttgart/Göttingen: B. G. Teubner Verlagsgesellschaft/Vandenhoeck & Ruprecht. 27쪽

6) Dilthey, W., GS VII 278쪽.

der geschichtlichen Welt in den Geisteswissenschaften)이다. 한편 이곳에서 기술되고 있는 '역사이성비판' 개념의 전모는 이보다 30년 정도 앞선 1883년『정신과학입문』(Einleitung in die Geisteswissenschaften)에서 이미 확연히 드러난다. 그는 이곳에서 파편화된 개별 학문들의 인식론적 토대의 필요성을 역설한다. 그러나『정식과학입문』집필을 포함한 '역사이성비판' 수립의 전체적인 계획은 이미 1883년 소위 "알토프 편지(Althoff-Brief)에서 구체적으로 언급된다. 이곳에서 그는 "모든 철학은 경험과학"이며 자아와 실재성(Wirklichkeit)의 대결구도를 표상하고 있는 경험론과 합리론의 싸움은 "영혼의 삶의 총체성"(Totalität des Seelenlebens)의 관점에서 접근할 때 해결될 수 있다고 주장하면서 그의 위에서 언급한 집필 계획을 펼쳐놓는다. 이렇듯 '역사이성비판' 기획은 그의 철학적 여정 곳곳에 다양한 사유의 활동들을 다양한 형태로 자리하고 있음에도 불구하고, 그의 학문적 생애를 관통하는 대전제라고 할 수 있다. 이러한 이유에서『정신과학의 역사적 세계의 건립』의 편집자는 '역사이성비판'이 그의 저술 전체를 묶는 이름이 될 수도 있다고 말한다.[7]

그는 '역사이성비판'을 "인간 자신과 '인간 자신이 형성해 낸 사회와 역사'를 아울러 인식하는 인간의 능력에 대한 비판"[8]이라 규정한다. 이 규정은 앞에서 언급한 이성의 자기비판, 이성을 통한 당대 학문의 시대정신에 대한 반성적 비판을 과제로 하는 칸트의 '이성비판'의 정신을 함축한다. '인간 자신을 인식하는 인간의 능력에 대한 비판은 곧 앎의 정초놓기(Grundlengung des Wissens)[9], "정신과학의 인식 이론적 토대 확장"[10]이라는 역사이성비판의 과제를 뜻한다. 이는 다름 아닌 칸트의 선험철학이 감행했었던 앎의 조건 물음과 관련된다. 이러한 의미의 '비판'

7) W. Dilthey, GS, VII, V.
8) W. Dilthey, GS, I, 116쪽.
9) W. Dilthey, GS, VII, 7쪽.
10) W. Dilthey, GS, I, 116쪽.

은 분석과 이를 바탕으로 한 구조화 작업이다. 앎의 토대 물음이라는 이성비판의 첫 번째 과제는 독일관념론의 전통에 따라 학문의 토대 물음으로 귀결되어 다시금 형이상학의 본질에 대한 물음으로 환원된다. 여기서 시대비판이라는 이성비판의 두 번째 계기가 형성된다. 건축술적 구조화라는 순수이성비판의 방법론이 역사이성비판의 시작점을 형성한 데 비해 자연과학을 전형(Vorbild)으로 삼은 형이상학의 재구조화라는 그 방법론의 기초이념은, 칸트가 그의 선배 철학자들에게 그러했듯, 딜타이의 비판의 시선에도 포착된다. 칸트를 자연 형이상학자로 규정한 딜타이는 그의 "철학적 사고는 뉴턴의 형태에 의해 각인된 것과 같은 수학적 자연과학들의 패권을 통하여 규정된"[11] 것이라 말한다. 수학이라는 공용어를 무기로 학문 세계의 지평 확대 전쟁을 감행하는 자연과학의 위협에 대항하고자 정신을 대상으로 하는 개별 학문들을 정신과학으로 총칭하여 '체험', '표현', '이해'를 그것의 근거[12]로 규정한 딜타이에게 있어 이러한 칸트의 방법론은 한편으로는 형이상학을 안전한 학문의 반석에 올려놓고자 하는 동의할만한 시도일 수 있지만, 다른 한편으로는 그것의 본질을 위협할 수 있는 위험한 문호개방의 시도였다. 뉴튼적 절대 시간이 내재화된 칸트의 선험철학적 시간관은 물자체-현상 구분의 시발점인 동시에 인식, 의식, 그리고 경험의 종착지이다. 외적 자연의 대상의 내재화(Innewerden)를 통한 생생한 인식의 조건 탐구인 역사이성비판의 측면에서 볼 때, 선험철학적 시간은 객관을 주관화할 수는 있어도 객관과 주관의 실제적(wirklich) 일치에 도달할 수는 없다. 딜타이에게 있어 시간은 직관의 형식이 아닌, "삶에 대한 최초의 범주적 규정"[13]이다. 그렇기에 그는 "로크와 흄, 그리고 칸트가 구성한 인식 주체의 혈관을 흐르는 것

11) 최성환. (2001).「칸트와 해석학」,『칸트와 현대유럽철학』, 철학과 현실사, 177-205쪽, 188쪽.

12) W. Dilthey, GS, VII, 131쪽

13) W. Dilthey, GS VII, 131쪽

은 진정한 피가 아니다. 그것은 그저 단순한 사유 활동으로서의 이성에서 희석된 묽은 즙에 불과하다"[14]고 말한다. 이렇듯 칸트에게서 배운 비판 정신으로 또 다시 칸트를 비판한 딜타이는 나아가 당시 학계에 편만한 실증주의를 비판한다.[15] 그는 프랑스 실증주의와 영국의 경험주의를 꼭 집어 비판하면서, "콩트와 실증주의자들, 밀과 경험주의자들은 역사적 현실성을 자연과학의 개념들과 방법들에 끼워 맞추기 위해서 이를 분절화시켜 버린 듯이 보인다"[16]고 말하면서 인문학의 자연 과학화를 비판한다.

'이성비판'의 유산은 딜타이와 약 반세기를 공유한 프랑크푸르트 학파의 비판 이론가 호르크하이머(Horkheimer)에게로 이어졌다. 일찍이 칸트 연구가였던 그[17]는 인간성 소외를 초래한 자본주의와 물질문명 팽창기에 이성의 객관적 사용의 복권을 역설한 『도구적 이성비판』을 발표하기 이전, 딜타이가 칸트적 비판 정신으로 칸트 자신을 자신의 비판의 범주에 복속시켰던 것과 유사하게 딜타이 비판으로부터 그의 이론을 전개해 나간다. 조금 부풀려 말하자면, 그가 볼 때 딜타이는 여전히 칸트주의자이

14) W. Dilthey, GS I, XVIII.

15) 이와 관련하여 최성환(2001)은 그가 정신과학의 토대를 마련하기 위한 "객관주의에의 추구가 실증주의를 의미하는 것은 아니다"(201, 202쪽)라고 말한다.

16) W. Dilthey, GS I, XVI. 그의 비판은 당시 역사학파에 집중되어 있다. 개별과학들은 18세기 직전까지 중세적 형이상학의 속박에 놓여 있었다. 그러나 18세기에 역사학을 위시한 개별학문들을 가두는 속박의 주체는 자연과학적 방법론으로 대체되었다. 예컨대 버클(H. Th. Buckle)은 그의 "영국 시민화의 역사"(History of Civilisation in England)에서 "자연과학의 원리들과 방법들을 전용함으로써 새롭게 역사적 세계의 수수께끼를 풀어보려는" 시도는 "자연과학적 방법론을 역사연구의 영역에 독단적으로 덧입힌 것"이다. W. Dilthey, Texte zur Kritik der historischen Vernunft(Hg. Hans-Ulrich Lessing), Vandenhoeck & Ruprecht, 11-12쪽 참조.

17) 그는 1922년 "Zur Antinomie der teleologischen Urteilskraft"라는 제목으로 박사학위 논문을 발표하였으며 연이어 1925년 교수자격취득논문으로 "Über Kants Kritik der Urteilskraft als Bindeglied zwischen theoretischer und prktischer Philosophie"을 발표하였다.

다. 그에 따르면 “칸트가 수학적 자연과학 안에서 인식 주체를 구현했던 것과 같이”, 딜타이가 구현하고자 했던 역사적 존재로서의 인간은 갈등을 배제한 채 전개되는 비현실적 역사성을 범주로 한다.[18] 딜타이의 역사이성비판은 우리의 삶이 펼쳐지는 구체적인 역사가 아닌, 칸트의 초월철학(transzendentale Philosophie)과 크게 다르지 않게, “우리 자신을 인식하기 위한 것”[19]이다. 이러한 관점에서 그는 이성비판의 이념을 현실화시킨다. 칸트의 이성비판이 이성에 의한 자기비판의 면모가 강했다면, 딜타이에게 이는 이성이 감행하는 사회 비판의 토대로서의 학문 세계 비판의 면모가 강했다. 호르크하이머는 이러한 이성비판의 토대를 확장적으로 계승하여 이성이 감행하는 사회 비판으로 곧장 나아간다.

그는 주관적 이성과 객관적 이성을 대립시키면서 이성이 도구화되고 있는 당시의 의식 구조를 비판하였다. 로고스로 상징되는 토대주의, 그리고 토미즘과 데카르트주의로 대표되는 실체 형이상학은 보편적 일자에 대한 희망을 전제로 펼쳐지는 객관적 이성의 전통적인 활동무대이다. 호르크하이머에 따르면 개별적 이성의 사적 사용을 넘어 개별 현상들의 본질적 공통분모를 찾고자 하는 철학의 노력은 자연과 사회에 내재되어 있는 객관적 이성을 신뢰하는 이성의 공적 사용이다. 반면 이성의 주관적 사용은 현실에 대한 이성의 순응을 나타낸다. 쉽게 말하자면 주관적 이성은 계산 도구로서의 이성을 가리킨다. 이는 “범주화, 추론 및 연역의 능력”[20]으로서 “목표에 도달하기 위한 절차적 방법”[21]과 관련된다. 이 모든 것은 자기보존이라는 이성적 생명체의 본능적 관심에서 비롯된 것이

18) 전석환. (2015).「비판이론에 있어서 딜타이비판 - 막스 호르크하이머(Max Horkheimer)의 초기사상을 중심으로」,『철학사상문화』, 20, 107쪽.

19) Horkheimer, M. (1988). Psychologie und Soziologie im Werk Wilhelm Diltheys, Gesammelte Schriften(이하 GS) Bd. IV, Fischer, 356쪽

20) Horkheimer, M. (1991). Zur Kritik der instrumentellen Vernunft, GS Bd. VI, Fischer, 27쪽

21) Horkheimer, M., GS Bd. VI, 27쪽

다. 그에 따르면 오늘날 이성은 이러한 본능적 관심에만 충실한 나머지, 체계와 통합, 진리와 본질을 희구하는 객관적 사용 능력을 스스로 축소시켰다. "단지 결과를 창출하는 수단에만 관계된 능력"[22]으로 도구화되어 버린 이성은 자율성을 상실한다. 그 결과 "사변적 이성은 죽음"[23]을 맞이한다. 역사이성비판이 이 죽음의 원인으로 '인문학의 자연과학화'를 지목한 것과 마찬가지로 도구적 이성비판은 실증주의를 꼽는다. 사유경제학(Denkökonomie)이 행하는 복잡한 논리적 작업은 "수학과 논리학적 기호의 근거가 되는 정신적 활동이 실제로 수행되는 것을 전적으로 배제한 채"[24] 진행된다. 사유경제학적인 입장에서 보면 "정의와 자유가 그 자체로 불의와 억압보다 더 좋은 것이라는 진술은 학문적으로 검증될 수 없기 때문에 무용하다는 진술은 빨간색이 파란색보다 더 아름답다"[25]는 진술과 마찬가지로 의미 없는 것이다. 이처럼 "실증주의자들은 철학을 과학에 순응시킨다."[26] 다음의 그의 언명은 디지털 원주민이 성장하고 있는 오늘날 우리의 도시에도 적중한다.

> "오늘날 사람들은 기본 개념들이 물리학과 기술의 전진을 통해 명백해진다는 환상에 현혹되어서 복잡성을 벗어던지려는 유혹에 너무나도 쉽게 빠져버린다. 산업주의는 심지어 철학자들에게까지 그들의 작품을 표준화된 식사 도구를 생산하는 과정으로 이해하도록 압력을 가한다." 이러한 압력은 "정신적인 욕구를 소책자 형태에 맞추어 재단하려는 인간적인 충동"에 기인한다.[27]

이러한 시대진단은 주관적 이성에 대한 객관적 이성의 우위 확보의 당위

22) Horkheimer, M., GS Bd. VI, 41쪽
23) Horkheimer, M., GS Bd. VI, 40쪽
24) Horkheimer, M., GS Bd. VI, 44쪽
25) Horkheimer, M., GS Bd. VI, 44쪽
26) Horkheimer, M., GS Bd. VI, 76쪽
27) Horkheimer, M., GS Bd. VI, 168. 169쪽

성으로 이어진다. “오로지 비판의 길만이 열려있다는 칸트의 준칙은 현재의 상황에도 잘 들어맞는다”는 그의 고백은 지금도 여전히 유효하다.

앞에서 살펴본 바와 같이 ‘이성비판’의 철학은 칸트의 유산을 따라 이성의 자기비판이라는 ‘소극적 의미’와 타자 비판이라는 ‘적극적 의미’를 갖는다. 적극적 의미의 비판의 대상을 우리는 또다시 두 가지로 나누었다. 첫째는 이성이 놓인 당시의 상황에 펼쳐졌던 학문에 대한 비판이고, 둘째는 학문이 펼쳐진 시대, 나아가 시대정신에 대한 비판이다. 그리고 ‘이성비판’의 철학을 관통하는 방법론은 분석과 이를 토대로 한 방향 제시라고 할 수 있다. 이제 ‘인공이성비판’은 첫 번째 의미를 따라 앞 장에서 ‘인공지능 인문학’이라는 주제로 묶일 수 있는 학문 분야들을 분석한다. 그리고 두 번째 의미를 따라 이 학문 분야들을 비판한다. ‘인공이성비판’이 시도하는 비판의 관점 역시 ‘이성비판’의 맥락을 그대로 이어받아 ‘체계에 대한 추구’, ‘이성의 자율성 회복’으로 삼는다. 이 두 과제는 칸트의 비판 철학의 체계에 맞추어 각각 ‘분석론’과 ‘변증론’이라는 이름 아래서 수행될 것이다.

3. 인공이성비판의 분석론: 인공지능 인문학의 범주

‘인공지능’을 주제로 한 인문학 연구, 즉 인공지능 인문학은 1990년대 이래로 한동안 잠잠하다가 최근 다시 유행처럼 번지고 있다. 알파고 이후 우리나라에서는 ‘인공지능’이 한국학술지인용정보(KCI) 인문학 관련 상위 키워드에 수년 동안 등재되어 있었고 관련 논문 수도 매년 증가하고 있다.[28] 『인공지능인문학연구』 등 이를 다루는 전문학술지도 생겨났다. 외국의 예를 들자면 국제적인 유수 출판 기업인 Springer는 『Minds

28) 이와 관련해서 “김형주. (2018). 「인공지능 철학 국내연구 동향 분석—인공지능 철학의 생장점에서—」. 『인공지능인문학연구』, 1, 149-170쪽”이 부분적인 이해에 도움이 될 수 있다.

and Machines』, 『AI and Ethics』와 같은 전문잡지를 수년 전부터 발간하여 관련된 담론 생성의 장을 꾸준히 넓히고 있다. 한편, 새로운 인문 정신으로 각광받는 포스트휴머니즘은 이러한 유행을 이끄는 선두주자 역할을 하고 있다. 이치구로가 2017년 노벨문학상을 수상한 이래 그의 『Never let me go』에 대한 포스트휴머니즘적 해석은 작년 문학계를 달구었다. 인공지능 인문학을 인공지능으로 빗어지는 인간을 둘러싼 여러 문화적, 사회적 현상들에 대한 인문학, 즉 인공지능을 소재로 하는 인문학으로 이해한다면,[29] 그리고 인공지능이 사회적 화두라는 사실을 생각한다면, 인공지능 인문학은 말 그대로 지금 시대의 인문학이다.

앞 장에서 언급했듯 이성비판의 첫째 소임, 즉 소극적 의미는 이성의 자기비판이다. 자기 비판은 자기를 구성하는 요소들에 대한 분석을 의미하고 이는 결과적으로 체계에 대한 추구로 이어진다. 주지하듯 '이성비판학'의 '비판'은 분리함을 의미하는 그리스어 크리노(κρίνω)를 어원으로 갖고 칸트도 이러한 어원학적 유산을 따른다.[30] 이러한 의미에서 이성비판의 첫째 의미는 이성의 능력들에 대한 분석, 즉 범주화라고 할 수 있다. 그가 인간의 인식능력을 감성, 지성, 상상력, 추론능력으로서의 이성 등으로 나눈 것도 이러한 이유에 기인한다. 인공이성비판은 여기에 시대비판이라는 두 번째 의미를 더하여 내재적 의미로서의 이성의 능력 비판이 아닌 현 시대의 이성의 활동에 대한 비판, 다시 말해 분석을 시도한다. 이러한 배경에서 인공이성비판은 현재 행해지고 있는 인공지능에 대한 인문 담론을 크게 네 가지 줄기로 분석한다.[31]

---

29) 김형주 & 이찬규. (2019).「포스트휴머니즘의 저편: 인공지능인문학 개념 정립을 위한 시론」.『철학탐구』, 53, 51-80쪽 참조

30) Paul Guyer. (2015). Kritik, Kant-Lexikon(Online), Berlin/New York: Walter de Gruyter, 1303쪽 참조

31) 후술하겠지만 이 네 가지는 인공지능 철학, 포스트휴머니즘, 디지털인문학, 인공지능 윤리이다. 이 구분은 논자의 연구 경험에 의거한 귀납적인 구분이기 때문에 이 네 영역간의 연역적인 체계를 설명하는 것은 어렵다. 한편 당연한 말일

첫째, 몸과 마음의 문제, 의식의 문제를 경험주의적 관점에서 조망하는 영미 심리철학을 배경으로 태동한 인공지능 철학(Philosophy of Artificial Intelligence)이다. 이는 인공지능의 인식론적 쟁점과 깊은 관련이 있다고 할 수 있다. 이러한 사실은 인공지능 철학이 인공지능 공학과 그 태동기를 함께 한다는 역사적 사실로부터 방증 될 수 있다. 20세기 초반과 중반, 철학은 독일을 위시한 유럽 중심의 형이상학적 관념론에 대한 회의와 더불어 명제논리중심의 언어철학이 득세하였다. 그 중심에는 반 심리주의(Anti-psychologism)를 주장한 프레게(Frege)를 선조로 그에게 직접적으로 영향을 받은 세 거목들인 러셀(Russell), 비트겐슈타인(Wittgenstein) 그리고 카르납(Carnap)으로 대표되는 논리실증주의라는 강력한 철학사조가 있었다.[32] 이들은 '의미는 진리에 선행한다'는 기조로 칸트가 인식가능의 영역과 불가능의 영역을 구분했듯, 인식적으로 유의미한 세계와 무의미한(meaningless) 세계를 구분하였다. 특히 그들은 2,000여 년 동안 철학의 고유 영역으로 여겨져 왔던 형이상학을 인식적으로 무의미한 사이비과학으로, 윤리학을 참, 거짓을 논할 수 있는 인식의 영역이 아닌 단지 정의주의(emotivism)로 칭하면서 학문의 영역에서 배제한다. 이렇듯, 그들은 어떤 대상영역이 의미를 지니려면 그것은 모두 언어로 표현이 가

수도 있겠지만, 이 네 영역의 외연이 명확하게 구분될 수도 없다. 예를 들어 디지털인문학의 방법론을 통하여 인공지능 철학과 관련한 결론을 도출할 수도 있다. 이렇듯 이 네 가지는 인공지능에 관한 인문학을 네 가지 주요점에 따라 구분한 것이라는 점에서 논리적 설명상의 한계를 갖는다고 할 수 있다. 그러나 나는 이 한계가 구분의 원천적인 불가능으로 이어진다고는 생각하지 않는다. 칸트도 선험적 종합판단의 필요조건으로 감성과 지성의 존재를 연역적으로 논증하지는 않았다. 나아가 그는 그의 비판철학의 결론이라고 할 수 있는 정언명령의 생성주체인 실천이성은 더 이상의 근거물음이 차단되어 있는 하나의 사실(Faktum)이라고 말한다. 이성의 요소들은 인식론적 그리고 도덕적 실재론에 근거하여 현상의 설명을 위해 추후적으로 요청되는 것이다. 칸트의 비판철학의 정신을 모델로 하는 인공이성비판의 이 구분도 용인될 수 있기를 희망한다.

32) Glymour, C., Ford, K. & Hayes, P. (2000). "The prehistory of android epistemology". *Artificial Intelligence: Critical Concepts, 1*, 113쪽

능해야 하고 검증 가능해야 한다고 주장한다.[33] 나아가 이를 위해서 언어체계의 정형(well-formed formula)을 형성하는 규칙들(formation rules)이 선재(先在)해야 함을 강하게 주장하고 실제로 인공언어라는 이름으로 이를 구성하기에 이른다.[34] 이러한 사상은 인공지능 창시자들의 사고를 지배해 온 철학적 기초가 되었고 이들이 "인공적인 프로그램 언어를 만드는 작업에도 논리실증주의의 영향은 결정적으로 작용했다."[35]

한편 시야를 확장하여 1950년 전후 당시 미국 철학의 지형도를 살펴보면 실용주의(Pragmatism)도 이와 더불어 큰 주류를 형성하고 있었음을 알 수 있다. 주지하듯 '문제해결'이라는 말은 인공지능을 정의할 때 빠지지 않고 등장한다.[36] 인공지능 교과서의 표준이라 할 수 있는 『Artificial Intelligence－A Modern Approach』에 따르면 인공지능은 지능이 필요한 문제를 해결해 주는 도구다. 실용주의자들은 인간이 마주한 사건들을 '문제'로 보고 이를 해결하는 과정을 삶의 과정으로 보았다. '문제해결'은 실용주의의 철학이었다. 이러한 철학은 인공지능의 역할을 규정하는데 큰 영향을 주었다.[37] 탐구 개념을 핵심으로 한 듀이의 문제해결에 관한 논의가 일반문제해결(General Problem Solving) 프로그램에 구현된 것이 대표적인 예라 할 수 있다. 또 실용주의의 창시자 퍼스(Pierce)가 문제해결의 방법으로 고안한 가설추리(abduction) 방식은 인공지능 예측프로그램에서 각광을 받고 있다는 사실도 주목할 만하다. 그는 대전제와 결론을

33) 관련하여 카르납은 '아름답다', '좋다'와 같은 것을 표현한 것은 은 실상 아무것도 한 것이다라고 말하며 형이상학적 명제의 무의미성을 역설한다. Carnap, R. (1931). *Überwindung der Metaphysik durch logische Analyse der Sprache.* Erkenntnis, 2, 219-241쪽. 236쪽 참조

34) 이초식. (1993).『인공지능의 철학』, 고려대학교 출판부, 71-75쪽 참조.

35) 이초식. (1993).『인공지능의 철학』, 고려대학교 출판부, 75쪽.

36) "인공지능 프로그램은 문제해결을 위해 고안되었다", Rich, E. (1987), *Encyclopedia of Artificial Intelligence: Artificial Intelligence.* John Wileys& Sons. 10쪽

37) Kieras, D. & Holyoak, K. (1987). Encyclopedia of Artificial Intelligence: Artificial Intelligence. John Wileys& Sons. 113-118쪽 참조.

토대로 소전제를 추론하는 가설추리(abduction)를 주창하였는데, 이것은 입력 값과 출력 값을 알려주고 함수를 최적화하는 심층신경망학습의 방법과 닮아있다.

1920년대 듀이(Dewey)가 실용주의의 대표자로 자리매김할 무렵, 논리실증주의가 미국 철학의 무대에 등장하였다. 이 두 진영의 교류는 점차로 활발해졌고 철학적인 공통분모가 구축되어 갔다. 이 둘은 진정한 지식의 원천은 경험이라는 것, 영국 경험론자를 자신들의 조상으로 여긴다는 것, 철학을 이론이 아니라 방법론으로 여긴다는 것 등을 공통분모로 갖는다.[38] 이러한 철학적 배경은 지능과 마음의 본질에 대한 탐구로 이어진다. 이는 튜링(Turing)의 "Computing Machinary and Intelligence"와 이에 대한 설(Searle)의 비판에서 비롯된 튜링테스트의 논쟁을 통해 본격적으로 주제화된다. 튜링이 소위 행동주의(Behaviorism)에 착안하여 지능적 존재자로서의 인공지능의 가능성 논제를 제시하였다면, 튜링은 의미론과 구문론의 구분을 통해 그 논제를 공격한다. 이 모든 논의는 철학을 학문의 방법론으로, 언어 분석을 철학의 본질로 여기는 영미 경험주의 철학을 배경으로 전개된다. 이러한 논의는 한편으로는 처칠랜드(Churchland)가 뇌 과학의 성과를 적극적으로 수용하여 주창한 신경철학(Neurophilosophy)으로, 다른 한편으로는 타가트(Tagard)의 전산과학철학(Computational Philosophy of Science), 플로리디(Floridi)의 정보 철학으로 이어지면서 인공지능 철학의 연구사를 형성하고 있다고 할 수 있다. 인공이성비판은 이 모든 흐름의 중심에 경험주의 철학이 자리하고 있다고 판단한다.

둘째, 포스트휴머니즘이다. 칸트로 대표되는 근대적 인간중심주의에 대한 회의에서 비롯된 이 이론은 문학, 철학, 역사 등 인문학 전반에 걸쳐 논의되고 있는 새로운 문화 운동이라 할 수 있다. 인공지능 철학이 지능적 존재자의 인식론적 물음에 천착했다면, 포스트휴머니즘은 기계를 포함

38) Nekrašas, E. (2001). "Pragmatism and positivism". *Problemos*, 59, 41-52쪽, 41쪽

한 모든 존재자들의 존재론적 지위 문제를 다룬다. 보스트롬(Bostrom)을 대표주자로 하여 인간 증강을 구호로 하는 트랜스휴머니즘까지도 확장적으로 포섭하는 이 이론은 시몽동(Simondon), 라투르(Latour), 플루셔(Flusser)등이 개진하고 있는 현대 기술·매체 철학에도 많은 영향을 주고 있다. 이 현대 담론의 근원은 서양의 근대에 닿아있다. "올리버 크뤼거(O. Krüger)는 포스트휴먼이라는 용어가 1656년 토마스 블런트의 글로소그래피아에 이미 등장하였다"[39]고 주장한다. 괴테의 『파우스트』에 등장하는 인공인간 호문쿨루스도 포스트휴먼적 상상력이 빚은 산물로 간주되기도 한다.[40]

한편 오늘날 우리가 규정하는 '포스트휴머니즘'에 부합하는 논의는 1900년대 후반 영미권 문학에서 본격적으로 발흥하였다. 미국의 문화이론가인 핫산(Hassan)은 "Prometheus as Performer: Toward a Posthumanist Culture?"라는 글을 1977년에 발표하면서 '포스트휴머니즘'이라는 용어를 학계에 다시 등장시켰다. 이 논문은 지금의 포스트휴머니즘 이론의 시작을 알리는 것으로 간주된다. 헤일즈(Hayles), 해러웨이(Haraway), 배드밍턴(Badmington), 울프(Wolfe)등이 대표적인 연구자로 꼽힌다. 앞에서 언급한 호문쿨루스가 포스트휴먼적 상상력의 원본 표상이라는 사실, 사이보그에 대한 이야기가 포스트휴먼 담론의 한 축을 형성하였다는 사실에서 짐작할 수 있듯, 포스트휴머니즘의 본질적인 문제의식은 인간과 비인간의 차별적 관계에 향해있다. 포스트휴먼 담론이 다양하게 분화되어 발달한 것은 사실이나, 이 이론들을 관통하는 공통의 토대는 "비 이분법 존재론(non-dualist ontology)"[41]이다. 비 이분법 존재론이라는 기조 아래 기존

39) 슈테판 헤어브레히터. (2012). 『포스트휴머니즘』, 김연순, 김응준 옮김. 성균관대학교 출판부. 54쪽.

40) 임석원. (2013). 「비판적 포스트휴머니즘의 기획: 배타적인 인간중심주의 극복」. 이화인문과학원 편, 『인간과 포스트휴머니즘』, 서울: 이화여자대학교출판부. 61-82쪽. 61쪽 참조.

41) Deretic, I, Sorgner, S (Eds.). (2016). *From Humanism to Meta-, Post- and*

의 논의를 바탕으로 포스트휴머니즘을 간략하게 정의하면 다음과 같다.

> "포스트휴머니즘은 모든 종류의 이분법은 해체되어야 할 폭력적 허상이라는 전제를 토대로 좁게는 육체와 정신의 이원론적 대립을 해체하고 "물질적인 것과 비물질적인 것, 동물과 인간 나아가 유기체와 기계"[42]의 경계를 허물어 "인간적인 것과 인간적이지 않은 것 사이의 이중대립"[43]을 해체한다. 이러한 입장에서 이는 바이오테크놀로지, 인공지능 등의 과학기술의 성과를 이용하여 생물학적 육체의 한계를 극복하는 포스트휴먼의 생성과 발전[44]을 추구한다."[45]

그 이름에서도 유추해 볼 수 있듯이 포스트휴머니즘은 포스트모더니즘이라는 사상적 토양에서 발전해 나왔다. 헤어브레이터(Herbrechter), 조르그너(Sorgner)와 같은 독일 문화권의 포스트휴머니즘 연구자들은 포스트모더니즘의 선구자 니체의 철학과 포스트휴머니즘의 연결 관계에 주목한다.[46] 조르그너에 따르면, 들뢰즈(Deleuze)와 푸코(Foucault)와 같은 영향력 있는 포스트모더니즘 연구자들의 생각은 포스트휴머니즘과 매우 밀접하게 연관되어 있다.[47] 그러나 인문학의 지형도를 보다 넓은 시각에서

Transhumanism?, Peter Lang. 14쪽.

42) Haraway, D. (1995). *in Wolfe*, 36. Quoted from Posthumanism, 190쪽

43) N. Badmington, Pool Alighty!; Or, Humanism, Posthumanism, and Strange Case of Invasion of the Body Snatchers, *Textual Practice 15*.1, 5-22쪽.

44) Kurzweil, R. (1999). "The Coming Merging of Mind and Machine-The accelerating pace of technological progress means that our intelligent creations will soon eclipse us--and that their creations will eventually eclipse". *Scientific American*, (2), 56-61쪽

45) 김형주 & 이찬규. (2019).「포스트휴머니즘의 저편: 인공지능인문학 개념 정립을 위한 시론」.『철학탐구』, 53, 51-80쪽, 57 이하.

46) 슈테판 헤어브레히터. (2012)『포스트휴머니즘』, 김연순, 김응준 옮김. 성균관대학교 출판부. 50-53쪽, 임석원. (2013), Sorgner, S. L. (2009). "Nietzsche, the overhuman, and transhumanism". Journal of Evolution and Technology, 20(1), 29-42쪽 참조.

바라보면, 포스트모더니즘이 전통 인문학에 대한 반작용에서 비롯되었으나 그 논의의 장은 언제나 인문학이었던 것과 유사하게 포스트휴머니즘이 전통 인문학과의 완전한 단절을 이루는 것은 사실상 어려운 듯이 보인다. 보스트롬은 "포스트휴머니즘은 완벽한 인간의 완전성, 이성성으로부터 직접 도출된 것이며 르네상스 휴머니즘과 계몽주의를 유산으로 물려받았다고 고백하면서" 근대정신의 완결 개념이라고도 할 수 있는 자율적 개인(autonomous individual)의 자율성(autonomy)이 그것의 핵심 가치(high value)인 반면, 인공지능 기술과 인터페이스 기술의 결합은 이를 위한 도구라고 말한다.[48] 인공이성비판은 포스트휴머니즘이 휴머니즘을 절대적 자아로 하여 포스트휴머니즘을 비아로 설정한다고 이해한다.[49]

셋째, 인공지능과 관련된 윤리적 문제를 다루는 인공지능 윤리(학)이다. 이는 이루다 사건, 마이크로 소프트의 테이 사건, 자율주행차 이슈 등 인공지능과 관련된 인공지능 시대의 생활세계와 밀접하게 관련되기 때문에 인공지능 철학, 포스트휴머니즘 담론과는 달리 학계를 넘어 산업계, 국가기관 및 국제기구에서도 관심을 갖는 주제라고 할 수 있다. 2021년에 조사한 바에 따르면 전 세계에는 최소한 84개의 인공지능 윤리강령이 존재한다. 대표적인 예로 가장 큰 국제 인공지능 협회인 IEEE는 『Ethically aligned Design』을 간행하면서 인공지능의 발전 방향에 대한 일종의 윤리적 기준을 제시하고자 노력하고 있다. 우리 정부도 '인간성'을 최고선으로 하고 '인간존엄성 원칙', '사회의 공공선 원칙', '기술의 합목적성 원칙' 3대 핵심 원칙으로 하는, "인간중심의 AI" 윤리 원칙을 발표하면서 이러한 세계적 흐름에 동참하고 있다.[50] 그리고 이를 수행하는 연구

47) Deretic, I, Sorgner, S (Eds.). (2016). *From Humanism to Meta-, Post- and Transhumanism?*, Peter Lang. 14쪽.

48) Bostrom, N. (2003). "The transhumanist FAQ". Readings in the Philosophy of Technology, 2, 355-360쪽.

49) Fichte, G.W. (1971). *GLW in: Fichtes Werk I*, Berlin: Walter du Gruyter. 20-26쪽 참조.

집단은 철학자들로 구성된 윤리학 연구 집단이라기보다는 인공지능 윤리 문제에 관심이 있거나 관련이 되어 있는 법학자, 공학자, 경영인, 행정인 등 다양한 배경의 사람으로 구성되어 있다. 연구자에게도 대중에게도 많이 노출된 '인공지능 윤리' 연구는 가이드라인 작성, 헌장과 원칙 제정을 위주로 하는 정책 연구의 성격이 강하다. 이러한 연구의 흐름은 긍정적인 평가와 함께, 이를 바라보는 입장에 따라 아쉬운 목소리를 듣기도 한다. 기업은 기술발전 저해를 우려하는 목소리를 내고 있고, 정치기구는 선언에 그친다는 비판을, 인문학계는 학문적 깊이의 부재를 걱정하고 있다.

넷째, 디지털 인문학이다. 빅 데이터와 컴퓨팅 파워를 무기로 전통적인 인문학적 방법과는 달리 정량적이고 거시적인 연구를 수행하는 디지털 인문학은 인문학 새로운 모습, 혹은 인문학 혁신의 모델로 회자된다.[51] 그런데 아직도 디지털 인문학의 주요주제 중 하나가 여전히 디지털 인문학의 정체성 확립 문제라는 사실은 아이러니하다. 그러나 이러한 아이러니함이 학제 간 융합을 지향하는 디지털 인문학의 학문적 성격을 드러내준다고도 할 수 있다. 이와 관련해서 학자들은 서로 다른 정의를 최소한 21개나 내놓았으며[52], 접근하는 관점에 따라 전산 인문학(Computing Humanities), 인문학 전산화(Humanities Computing)로 불리기도 하고[53], 중점을 두는 주제에 따라 공간 인문학(Spatial Humanities)로 불리기도 한

50) 김봉제 외 11인. (2020).「윤리적 인공지능을 위한 국가정책 수립」.『정책연구』, 2020(7), 1-235쪽 참조.

51) 스탠포드(Stanford) 대학의 디지털 인문학 관련 홈페이지(https://digitalhumanities.stanford.edu/projects)는 디지털 인문학의 학문적 성격에 대한 손쉬운 이해에 도움이 된다. 1890년과 1930년 사이에 발생한 유럽, 북미 인문학자, 종교인 250명의 서신 교환, 인용관계에 대한 정량적인 데이터 연구를 토대로, 그들 간의 영향관계를 시각적으로 제시하는 연구가 대표적이다.

52) Gardiner, E., & Musto, R. G. (2015). *The digital humanities: A primer for students and scholars*. Cambridge University Press. 3쪽.

53) 그러나 디지털 인문학과 인문학 전산화는 구분되어 논의되기도 하는데, 전자는 방법론을 후자는 학문분과를 의미한다. 같은 책 4쪽 참조.

다. “많은 인문학자들은 디지털 인문학을 인문학의 전통적인 작업을 위한 도구를 제공하는 방법론으로 보는 경향이 있는 데 비해, 컴퓨터 공학자는 디지털 인문학을 전자 양식(electronic form)이 이를 사용하는 학문분야에 어떠한 영향을 주는지, 이러한 학문분과들이 우리의 전산 지식에 어떠한 기여를 해야 하는지를 탐구하는 연구로 보는 경향이 있다.”[54] 이렇듯 디지털 인문학이라는 단일한 대상은 그것을 보는 자에 따라 다른 실체를 갖는다. 한편 스스로를 ‘디지털 인문학자’로 규정하는 사람들은 이 융합학문이 “정보통신기술의 도움을 받아 새로운 방식으로 수행하는 인문학 연구”로 “전통적인 인문학의 주제를 계승하면서 연구 방법 면에서 디지털 기술을 활용하는 연구”[55]라는 점에는 동의한다. 이 연구의 수행은 주로 사회과학연구에서 사용되는 디지털 툴(tool)을 활용한 인문학적 가치 평가가 요구되는 자료의 수집, 툴을 사용한 분석과 추론, (경우에 따라서는 이에 대한 연구자의 해석), 결과 값에 대한 데이터 시각화 순으로 이루어진다. 바로 이러한 과정 자체가 디지털 인문학에 대한 본질 규정이라고 할 수 있다. 즉 디지털 인문학이 무엇인지는 개념적으로 규정될 수 있는 것이 아니라 수행 과정을 통해 보여질 수만 있는 것이다.

4. 인공이성비판의 변증론: 인공지능 인문학 비판

지금까지의 논의를 이끌어 오고 있는 ‘이성비판’의 특징을 한 마디로 요약하면 ‘분석적 체계화에 따른 동 근원 밝힘’이라 할 수 있다. 이러한 이유에서 ‘인공이성비판’은 ‘인공지능’이 우리를 둘러싼 상황에 대한 분석적 체계화, 그리고 분석 결과들의 공통 토대를 밝히는 작업이라 할 수 있다. 한편 나는 칸트를 따라 비판을 소극적 의미와 적극적 의미로

54) Gardiner, E., & Musto, R. G. (2015). 3쪽.

55) 김현. (2013).「디지털 인문학: 인문학과 문화콘텐츠의 상생 구도에 관한 구상」. 『인문콘텐츠』, (29), 9-26쪽.

구분하여 논의를 전개하고 있는데 3장은 그 첫 번째 부분에 해당하는 논의였다. 이번 장은 이성비판의 시각에서 바라보는 인공지능 인문학에 대한 비판을 다룬다. 그리고 그 과정을 통해 '재귀적 인간중심주의'를 인공이성비판이 실행하는 비판의 관점으로 도출한다. 나는 이를 비판의 적극적 의미로 삼는다.

실증주의, 실용주의, 환원주의, 그리고 신경중심주의를 품고 전개되는 인공지능 철학은 앞에서 언급했듯 경험주의에 그 뿌리를 둔다. 인공지능이라는 기술에 대한 신뢰와 기대가 깊어질수록 경험주의는 한편으로는 형이상학에 대한 회의주의, 다른 한편으로는 유물론 절대주의 성향을 더해간다. 그것의 주요 논구 대상이 '지능'이기 때문에 이러한 경향은 비단 인식론에 국한되지 않고, 실천 철학의 영역에도 영향을 미친다. 유기체의 욕구, 욕망, 자유 의지에 대한 정량적 이해는 그 자체로는 아무런 문제가 없다. 오히려 보이지 않는 것에 대한 정량적 이해는 좋은 이해를 위한 초석이라 할 수 있다. 문제는 오로지 그것만이 전부라는 신념, 그 신념을 구현하기 위한 부단한 시도에 있다. 생각하기라는 비가시적 활동을 시각화하여 제시하는 것은 생각의 본질을 밝히기 위한 중요한 필요조건이지만 충분조건은 아니다. 인공이성비판은 생각의 결과는 뇌라는 신체의 움직임으로 이어지지만, 이 사실이 그 반대도 긍정하는 것은 아니라는 정신철학의 입장에 동의한다.[56] 약간의 과장을 보태면 물리주의적 경험주의는 호르몬-전기 환원주의를 넘어 사회적 물신주의의 탄탄한 전제 역할을 하기도 한다. 눈에 보이지 않는 비즉자적인 정신 가치가 물화되지 않을 때, 그럼에도 불구하고 그 존재의 당위성을 주장하는 것이 점차 옹색해지는 것은 과학주의가 인공지능 시대를 잠식하고 있는 세계관으로서의 영향력을 점차 확대시켜가고 있기 때문이다. 큰 유익을 낳는 첨단의 인공지능 기술인 딥 러닝(Deep Learning)에 대한 사회적 문제가 제기되는

56) Gabriel, M. (2015). *Ich ist nicht Gehirn: Philosophie des Geistes für das 21. Jahrhundert.* Ullstein. 18-23쪽 참조

이유는 무엇일까? 뇌 신경을 모방하여 만든 인공신경망 모델의 뒤에는 '개연성'에 의거한 문제해결을 최고의 가치로 삼는 실용주의적 사고가 자리하고 있다. 그리고 문제 해결 과정에 대한 과학적 설명력을 일정 부분 포기하고 생산성 향상에 집중하는 딥 러닝의 정신은 자본주의 정신과 시너지 효과를 낸다.[57] 이 모든 것의 뒤에는 모든 지식과 자극, 정보는 양화가 가능하다는 물리적 환원주의가 자리하고 있다.

한편 포스트휴머니즘도 이러한 사고체계를 근본적으로 공유한다. 인공이성비판은 앞에서 밝혔듯 포스트휴머니즘적 사고의 궁극적인 목표는 모든 이분법의 해체라고 이해한다. 인공지능을 포함한 비인간 존재자와 인간 사이의 수평적 존재 위계를 피력하는 라투르의 행위자 연결망 이론(Actor-Network Theory), 코드화된 사고로 점철된 디지털 사회의 커뮤니케이션의 새로운 방향성을 제시하고 체계화한 플루셔의 코무니콜로기(Kommunikologie)등은 정보와 기술, 나아가 인공지능의 등장으로 인해 변화되는 세계를 그에 알맞은 새로운 시각으로 분석한다. 이 이론들이 각광받는 공통된 이유 중 한 가지는 하나같이 탈근대라는 기치를 든다는 것이다. 요컨대 지금과 같은 인공지능 시대는 기존과 같은 인간중심주의로는 메울 수 없는 해석적 공백이 발생하기 때문에 미래에 알맞은 철학

57) 인공지능은 인간의 뇌의 기능 모방에 착안한 기호주의(symbolism)와 그것의 작동원리를 도식화한 것에 힌트를 얻어 구성된 연결주의 진영으로 양분되어 발전되어 왔다는 것은 정설로 여겨진다. 또한 기호주의, 연결주의가 갖는 각각의 특징을 규칙과 학습이므로 이 양자는 규칙기반 인공지능과 학습기반 인공지능으로 구분되기도 한다. 거칠게 말하자면 전자가 연역논리에 기반한다면 후자는 귀납논리에 기반한다. 전자에는 대표적으로 전문가 시스템, 결정나무 모델들이 속하고 후자에는 2006년 힌튼(J. Hinton)이 개발한 심층기계학습(Deep Learning) 기법이 속한다. 전자는 설명력이 높은 반면, 효율성이 떨어지고, 후자는 설명력이 낮은 반면 효율성이 높다. 인공지능 기술에 대한 자세한 설명은 본고의 목적에서 벗어나므로 생략하기로 한다. 이와 관련한 더 자세한 논의는 졸고 2022년 4월 출간 예정인 Kant and Artificial Intelligence(Ed. Hyeongjoo Kim, Deiter Schönecker. De Gruyter)의 4장 Tracing the Origins of Artificial Intelligence: A Kantian Responseto McCarthy's Call for Philosophical Help에서 다루었다.

은 이분법의 형이상학적 전제의 부당성을 폭로하고 새로운 전제 위에서 새로운 가치를 지향해야 한다는 것이 이상에서 언급한 사조의 공통분모다. 이분법을 해체하는 방법은 크게 두 종류로 구분된다. 분리된 A와 B를 초월한 C를 상정하여 A와 B를 C에 복속시키는 일원론적 전략이 그 하나라면, A를 B로 환원시키던지, 아니면 B를 A로 환원시키는 전략도 있다. 포스트휴머니즘은 이 두 전략을 적절히 이용한다. 인간과 기계, 달리 말하면 인격과 인공지능의 차이를 불식시키기 위해, 지능 개념의 보편화를 시도한다. '지능'의 본질은 문제해결을 위한 계산능력이고 이 능력은 인간뿐 아니라 동물, 컴퓨터 등이 공유하고 있다는 인공지능 공학의 근본적인 입장을 받아들인다. 그 후 이렇게 정의한 지능은 휴머니즘과의 대결 구도를 통해, 다시 기계 지능의 본질적인 속성으로 재 위치된다. 휴머니즘은 이렇듯 협소한 지능의 능력이 인간 지능의 전부라 결단코 이야기하지 않는 반면, 포스트휴머니즘은 이러한 인식론적 논의는 존재론적 논의로 확장되어 인간 존재의 재규정이 이루어진다. 미래라는 미지의 세계에 자신을 던지는 현존재의 자기구성력, 자기 자신을 그러한 능력을 지닌 존재로 인식하는 자기 인식력 등은 모두 계산능력의 변용으로 환원되거나, 아니면 허상으로 치부되어 비인간존재자와의 존재론적 동일성은 정당화된다. 이에 인공이성비판은 다음과 같이 묻는다. '건조한 계산 덩어리인 인간이 미래의 인간, 즉 포스트휴먼인가?' 마르쿠스 가브리엘의 『나는 뇌가 아니다』의 다음의 구절은 이 물음에 대한 적실한 답을 제시한다.

> "건강한 뇌가 없다면, 우리[인간: 김형주]가 존재할 수 없다는 것은 당연하다. 우리는 생각할 수도 없고, 깨어있을 수도, 의식을 가지고 살 수도 없다. 그러나 이 사실로부터, 만약 우리가 많은 부가적인 논증을 덧대지 않는다면, 우리가 우리 뇌와 동일하다는 사실이 도출되지는 않는다."[58]

비가시적 활동인 사고활동의 필요조건인 가시적인 물질 덩어리 뇌는 사고활동의 필요조건이지 충분조건이 아니라는 비교적 상식적인 그의 주장을 받아들인다면, 새로운 인문학을 자신하는 포스트휴머니즘의 주장의 전부를 긍정하긴 어려울 것이다. 인공이성비판은 신경중심철학이 정신철학(Philosophie des Geist)의 하위 분과이라는 그의 주장[59]에 동의한다.

인공지능 철학과 포스트휴머니즘의 본질에는 비가시적인 것들의 가시화 시도와 그렇게 될 수 있다는 믿음이 놓여있다. 이러한 믿음이 가장 가시적으로 드러나는 영역이 디지털인문학이다. '인문학을 위한 방법론이 인문학이 될 수 있는가?'라는 질문은 디지털인문학의 정체성과 관련하여 줄곧 제기되어 왔다. 이 질문은 '어떤 방법론의 대상이 인문학 자료이면, 그것도 내용학으로서 인문학이 될 수 있는가?'라는 질문을 함축한다. 만약 이 질문에 대한 답이 긍정이라면 방법론으로서 디지털 사회학, 방법론으로서 디지털 예술, 심지어 디지털학 그 자체인 전산학과 방법론으로서의 디지털 인문학의 차이는 없다. 다만 각각의 방법론이 대상으로 하는 내용만 상이할 뿐이다. 그렇다면 디지털 인문학의 학문적 논쟁은 쉽게 종결될 수 있어 보인다. 앞에서 대다수의 인문학자들의 관점을 빌려 이야기했듯, 디지털 인문학은 인문학을 위한 수단일 뿐, 인문학 그 자체는 아니라고 하는 다소 보수적이지만 논리적인 결론은 학문성 논쟁의 종지부를 찍을 수 있는 듯이 보인다. 인공이성비판은 이 상식적인 결론에 굳이 반대하지 않는다. 특히 철학이 철학함을 통해 스스로를 전개시키는 방법과 비교해 보면 그 차이는 명확히 드러난다. 그러나 철학이 다른 개별학문과 맺는 관계와 디지털 인문학이 전통 인문학과 맺는 관계를 유비시켜 보면, 우리가 디지털 인문학에 부여한 생경함은 어느 정도 상쇄될 수 있다. 학문 생태계의 일원으로서 철학의 중요한 역할 중 하나는 학문

58) Gabriel, M. (2015). 41, 42쪽
59) Gabriel, M. (2015). 12쪽 참조

방법론의 제시이다. 이는 체계 구성과 본질 탐구라는 역할로 드러난다. 예컨대 법에 대한 본질적 탐구는 법철학이고, 과학의 본질에 대한 탐구는 과학철학으로 일컬어진다. 수리철학, 언어철학, 정치철학 등도 마찬가지다. 이처럼 철학이 개별학문과 맺는 관계는 디지털 인문학이 인문학과 맺는 관계와 유사하다.

인공이성비판은 좀 더 본질적인 부분에 주목한다. '부분에서 전체로, 전체에서 부분으로'라는 해석학적 순환의 표어는 인문학적 지식이 습득되는 과정의 핵심사항을 압축적으로 드러낸다. 주지하듯 전체를 이해하기 위해서는 부분들에 대한 이해가 요구되고, 부분들을 이해하기 위해서는 전체에 대한 선이해가 요구된다. 나아가 이러한 순환의 과정이 반복될수록 대상으로 삼은 텍스트에 대한 이해는 깊어지고 이와 관련된 새로운 의미들이 생성되어, 다른 차원의 의미 지평이 열리게 된다. 한편 디지털 인문학은 대상 전체의 파악과 양적 정보에 집중한다. 인간의 힘으로는 담을 수 없는 방대한 데이터를 최대한 많이 확보하고 그 양이 나타내는 의미를 제시해보는 것이 이 새로운 인문학의 목표라고 할 수 있다. 자료 전체에 대한 통합적인 이해가 선행해야 유의미한 연구가 될 수 있다는 전제는 그 자체로 옳다. 하지만 인문학적 의미는 순환의 곱씹음을 통해 생성된다는 것도 간과할 수 없는 중요한 사실이다. 디지털 인문학의 표어는 '부분에서 전체로'인 듯하다. 데이터의 양적 집적이 의미의 질적 고양으로 곧장 이어지는 것은 아니다. 질적 고양은 반성과 순환에서 일어나며 그 순환의 동력과 결과도 보태는 말없이 인문학이라는 이름 자체에 이미 전제되어 있는 인문 정신이다. 순진하게 무딘 벤담의 칼날은 다채롭고 섬세한 밀의 칼솜씨를 따라가고자 노력하지 않는다.

마지막으로 인공지능 시대의 인간의 존재론적, 윤리적 지위를 논의하고, 논의의 결과를 토대로 실천적 지침을 제시하는 인공지능 윤리에 대한 인공이성비판의 입장을 살펴보겠다. 인공이성비판은 포스트휴머니즘이 이미 낡아 폐기해야 할 것으로 여기는 근대성에 다시 주목한다. 근대성

의 윤리적 귀결은 인간중심주의(anthropocentrism)라 할 수 있다. 오직 인간만이 고유한(intrinsic) 가치를 가진다는 인간중심주의[60]는 표현상 인공지능 시대가 표방하는 윤리관에도 어색하지 않게 어우러진다. 2017년 EU는 인공적 존재자에게 법인격을 부여하자는 의견을 내놓았다. 우리나라도 인공지능에게 법인격을 부여하는 법안을 상정하기도 하였다. 당시로서는 파격적이었던 이 주장으로부터 우리는 인공적 존재자의 가치가 한 층 고양되었음을 감지할 수 있었다. 이 무렵 초지능, 특이점, 강 인공지능 등의 말이 함께 유행하였고, 아울러 '로봇과의 공존'과 같은 표어들은 우리 인간의 새로운 삶의 태도와 소통의 자세를 규정해 주었다. 그러나 지금은 앞에서 언급한 우리 정부의 발언에서도 알 수 있듯이 '인간 중심(Human-centered) AI'라는 표현이 세계 각국, 각 기관의 표어 선택권을 장악하고 있다고 해도 과언이 아니다. 'anthro'와 'human'이 각각 그리스어와 라틴어에 어원을 두고 있다는 표현상의 차이만 있을 뿐, 양자가 동의어라는 사실을 생각하면 지금 AI 시대가 요구하는 존재 이해에 '인간중심주의'가 중요한 자리를 차지한다고 생각할 수 있다. 그러나 이 근대를 대표하는 핵심 사상은 포스트휴머니즘의 맹공이 있기 전부터 환경주의, 생태주의, 동물해방론 등으로부터 이미 많은 비판에 충분한 시간 동안 노출되어 있었다. 더욱이 비 이분법 존재론이 정설인 듯 보이는 포스트휴먼 시대에 인간중심주의는 시대착오적 발상으로 간주된다. 그러나 한 번 더 생각해보면 지성으로 습득한 세계 이해와 존재의 심연에 뿌리를 둔 자기 이해의 괴리는 인간중심주의의 건재를 꾸준히 증명한다. 인공이성비판이 권하는 이성의 자기반성은 우리를 종국에는 인간중심주의로 데려간다.

인간중심주의는 크게 세 가지 층 차로 나누어진다. 가장 일차적이고 오랜 역사를 지닌 형이상학적 인간중심주의는 인간이 세계 내 모든 존재

60) Callicott, J. B., & Frodeman, R. (2009). *Encyclopedia of environmental ethics and philosophy (Vol. 1, pp. 223-225)*. Macmillan reference USA. 58쪽.

자들의 가치 위계에서 으뜸가는 위치를 갖는다는 입장을 피력한다. 태초에 신이 인간에게 자기의 숨, 즉 정신과 영혼을 불어넣어 자기 형상대로 인간을 만들었기 때문에 인간은 고유한 내재적 가치를 지닌다는 기독교의 주장도 형이상학적 인간중심주의의 한 예라고 할 수 있다. 형이상학적 인간중심주의는 인간 존재 자체의 독존적인 가치를 정당화하고 설명하기 때문에 존재론적 인간중심주의라고도 불린다. 이는 둘째, 도덕적 인간중심주의의 이론적 근거 역할을 한다. 근대 시기에 본격적으로 논의된 도덕적 인간중심주의는 인간(종)만이 직접적인 윤리적 고려대상의 집단(base class)에 정당하게 속한다고 주장한다.[61] 이에 따르면 식물, 동물, 생명체 혹은 자연물도 간접적인 도덕적 고려대상일 수는 있어도 도덕적 사고와 행위 주체는 아니다. 인간 외의 존재자들은 인간에 의해 도덕성을 부여받을 뿐, 인간만이 고유한(intrinsic) 도덕적 가치를 지닌다. 그러나 이 입장은 그 근거 물음에 있어 형이상학적 인간중심주의로 회귀한다. 이를테면 도덕적 인간중심주의를 대표하는 칸트의 '목적 그 자체로서의 이성적 존재자'라는 표어는 그것이 비록 종교에 의존하지는 않고 있지만 '이성'이라는 비가시적 실체에 의존하고 있기 때문에 존재론적 인간 중심주의를 전제하고 있다고 할 수 있다. 그러나 '새로운 지능적 행위자'라는 말로 인공지능을 세계 내 존재자의 한 구성원으로 받아들여야 한다는 목소리가 점점 힘을 얻고 있는 지금, 과연 우리는 '이성'의 선험적 존재성을 이유로 더 이상의 근거 물음을 효율적으로 차단할 수 있는가? 이 질문에 마지막 셋째, 재귀적(tautological) 인간중심주의는 '아니오'로 답한다. 그럼에도 불구하고 이는 여전히 인간중심주의를 주장한다. 이에 따르면 "인간이 가치를 경험하는 모든 방법을 포함하는 인간의 모든 경험은 인간의 경험이다."[62] 분석적 인간중심주의, 인식론적 인간중심주의라고도

61) Baird Callicott, J. (2013), *Thinking Like a Planet,* Oxford University Press, 9쪽.
62) 같은 책, 10쪽.

불리는 이 입장에 따르면, 인간이 인간 중심주의적 입장을 견지할 수밖에 없는 이유는 인간이 다른 존재의 삶을 경험할 수 없기 때문이다. 가령 '인공지능은 도덕의 중심에 서 있다'라는 말로 어떤 인간이 인공지능 중심주의를 주장한다고 해보자. 이 문장은 그 인간은 '인공지능은 도덕의 중심에 서 있다고 생각한다'라는 주장과 다르지 않다. 인공지능 시대에 인간이 중심이 되어야한다, 보다 정확히 말해서 '인공지능 시대에 인간이 중심이다'는 말은 인공지능을 비롯한 비인간 존재자들에 대한 인간의 존재론적 우월성을 주장하지 않는다. 이는 자신의 인식론적 한계에 대한 겸허한 고백에 불과하다.

'재귀적 인간중심주의'는 근대적 세계관의 상징이자 '이성비판' 계보의 시금석이라 할 수 있는 칸트의 '선험적 관념론'(transzendentaler Idealismus)과 인간 이해의 궤를 같이한다. 이는 선험적 관념론이 인간의 자기 인식의 한계를 자신 안으로 귀속시키면서 겸손한 이성 사용을 통한 안전한 학문의 길을 주장한 것과 마찬가지로 인간이라는 한계적 존재는 자기의 관점에서 세계를 구성할 수밖에 없다는 사실을 시인한다. 다시 말해 인간과 인공지능을 포함한 비인간 존재의 존재론적 지위, 윤리적 지위의 비차등성은 이론적으로는 가능할지 모르지만, 인식론적 평등은 원천적으로 불가능하다. 그리고 수평적인 인식론적 지위 보장이 불가능하다면 이는 자연히 존재론적, 윤리적 지위의 무차별적 평등에 대한 회의주의로 이어지게 된다. 이렇듯 '재귀적 인간중심주의'는 '인간 종 중심주의'와 구별된다. 후자가 근대와 지금을 관통하는 세계관이라면 전자는 인공지능 시대를 살아가는 이성적 존재자로서의 인간의 솔직한 자기 시인이다. 포스트 휴먼적 사고가 탈근대를 표방하고, 인간 종 중심주의가 온전한 근대성을 상징한다면 '인공이성비판'의 기초이념인 '재귀적 인간중심주의'는 인공지능 시대를 이성의 눈으로 재단할 수 있는 새로운 근대성이다.

5. 나가며

나는 이 글에서 인공이성비판 기획의 가능성을 검토해 보기 위해 우선 칸트, 딜타이, 호르크하이머의 이성비판 기획을 차례로 검토하면서 이성비판 기획의 계보를 개략적으로 구성해 보았다. 이를 통해 근대적 계몽정신을 상징하는 칸트의 이성비판 개념이 시대를 달리하며 지속하면서 어떻게 자기 자신을 드러내고 있는지 펼쳐 보이고자 하였다. 칸트가 수립한 이성비판의 의미에는 이성의 자기비판으로서 이성을 구성하는 요소들에 대한 분석, 이성의 타자 비판으로서 당대의 학문 비판, 나아가 학문과 교통하고 있는 시대정신에 대한 비판이 속해 있다는 사실을 논구하였다. 그리고 앞에서 언급 한 세 철학자들은 강조점을 달리하여, 칸트에게서는 첫 번째, 딜타이에게서는 두 번째, 호르크하이머에게서는 세 번째 의미를 비교적 선명하게 발견할 수 있다는 사실을 밝혔다. 그리고 이들을 관통하는 근본정신은 '이성의 자율성 복권', '체계에 대한 추구'라는 사실을 밝혔다. 이를 토대로 지금 이성이 처한 시대는 종교의 시대, 자연과학주의의 시대, 기술-자본주의의 시대를 지나 인공지능 시대라고 진단하고 인공지능을 대하는 이성의 활동을 인공지능 인문학이라 이름하였다. 그리고 위에서 언급한 비판의 첫 번째 역할에 따라 인공지능 인문학을 인공지능 철학, 포스트휴머니즘, 디지털 인문학, 인공지능 윤리로 범주화하였다. 그리고 '체계에 대한 추구', '자율적 이성의 복권'이라는 '이성비판'의 시각에서 이렇게 범주화된 인공지능 인문학을 비판해 보면, 이들의 공통된 특성으로 경험주의적 양화주의(empiricist quantificationism)가 도출된다고 논증하였다. 그리고 지금이 아무리 과학 기술의 첨단을 달리는 인공지능 시대라고 할지라도 이러한 입장이 인문학을 포함한 모든 학문 단위에 천편일률적으로 적용되는 것은 바람직하지 않다고, 즉 이러한 입장이 시대 정신이 되는 것은 바람직하지 않다고 주장하였다. 인공이성비판은 방향설정의 학문으로서 철학의 역할과 이성 비판의 정신에 입각하

여 근대윤리학의 주된 관점인 인간중심주의에 다시 주목한다. 그리고 재귀적 인간중심주의야말로 인간의 실존적 존재 긍정과 인식론적 한계를 수평적으로 받아들인 이성의 솔직한 자기 고백이라 간주하여 이를 인공지능 윤리가 지향해야 할 새로운 기조라고 평가하였다.

이 글의 제목인 '인공이성비판의 가능성 물음'이 이미 밝히고 있듯, 나는 이 글은 단지 결핍된 완결성만을 갖고 있고, 그렇기 때문에 한계를 갖고 있다고 자평한다. 인공이성비판 수립이라는 장기 프로젝트의 첫 문장을 담은 이 글은 인공이성비판의 전체 지형도를 구상해보고 내용 요소들을 추려 보는 것, 이를 통해 연구의 가능성을 타진해 보는 것을 목표로 한다. 이 사실이 칸트, 딜타이, 호르크하이머와 같은 철학자들의 이론을 깊이 있게 다루지 못한 것에 대한 적당한 이유가 될 수 있을지도 모른다는 생각이 작은 위로를 주기도 한다. 그러나 이 안도감이 이들의 철학을 바탕으로 구성한 '이성비판' 계보학이 나타내는 선명한 이론적 공백에 따른 아쉬움마저 덮지는 못한다. 또한 나는 이성의 자기비판, 학문 비판, 시대비판을 인공이성비판의 역할로 제기하였다. 이 글에서는 처음 두 과제에 대한 시론적 연구만이 시도되었다. 나는 나와 글의 한계를 잘 알고 있다. 한계에 대한 명확한 인식은 후속 연구에 대한 명확한 동기가 된다.

참고문헌

국립국어원. (1999).『표준국어대사전』. 두산동아

김봉제 외 11인. (2020).「윤리적 인공지능을 위한 국가정책 수립」.『정책연구』, 2020(7), 1-235쪽

김현. (2013).「디지털 인문학: 인문학과 문화콘텐츠의 상생 구도에 관한 구상」.『인문콘텐츠』, (29), 9-26쪽

김형주 & 이찬규. (2019).「포스트휴머니즘의 저편: 인공지능인문학 개념 정립을 위한 시론」.『철학탐구』, 53, 51-80쪽

김형주. (2018).「인공지능 철학 국내연구 동향 분석—인공지능 철학의 생장점에서—」.『인공지능인문학연구』, 1, 149-170쪽

슈테판 헤어브레히터. (2012)『포스트휴머니즘』, 김연순, 김응준 옮김. 성균관대학교 출판부

이초식, (1993),『인공지능의 철학』, 고려대학교 출판부

임석원. (2013).「비판적 포스트휴머니즘의 기획: 배타적인 인간중심주의 극복」. 이화인문과학원 편,『인간과 포스트휴머니즘』, 서울: 이화여자대학교출판부. 61-82쪽

전석환. (2015).「비판이론에 있어서 딜타이비판 - 막스 호르크하이머(Max Horkheimer)의 초기사상을 중심으로.」,『철학사상문화』, 20, 107쪽

정대성. (2012).「비판이론에 나타난 '바판' 개념의 의미 연구」,『가톨릭철학』, 18, 131-157쪽

최성환. (2001).「칸트와 해석학」,『칸트와 현대유럽철학』, 철학과 현실사, 177-205쪽

호르크하이머. (2006). 『도구적 이성비판』, 박구용 옮김, 문예출판사

Badmington. N. (2001) "Pool Alighty!; Or, Humanism, Posthumanism, and Strange Case of Invasion of the Body Snatchers", *Textual Practice 15.1*, 5-22쪽

Baird Callicott, J. (2013), *Thinking Like a Planet,* Oxford University Press, 9쪽

Boden, M. (Hrsg.) (1990). *The Philosophy of Artificial Intelligence*. Oxford: Oxford University Press

Bostrom, N. (2003). "The transhumanist FAQ". *Readings in the Philosophy of Technology*, 2, 355-360쪽

Bostrom, N. (2005). "A history of transhumanist thought". *Journal of evolution and technology, 14*(1). Retrieverd from https://nickbostrom.com/papers/history.pdf

Callicott, J. B. & Frodeman, R. (2009). *Encyclopedia of environmental ethics and philosophy (Vol. 1, pp. 223-225)*. Macmillan reference USA

Carnap, R. (1931). *Überwindung der Metaphysik durch logische Analyse der Sprache*. Erkenntnis, 2, 219-241쪽

Deretic, I, Sorgner, S (Eds.). (2016). *From Humanism to Meta-, Post- and Transhumanism?*, Peter Lang. 14쪽.

Fichte, G.W. (1971). *GLW in: Fichtes Werk I*, Berlin: Walter du Gruyter

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI.* Berkman Klein Center Research Publication, (2020-1)

Gabriel, M. (2015). *Ich ist nicht Gehirn: Philosophie des Geistes für das 21. Jahrhundert.* Ullstein

Gardiner, E., & Musto, R. G. (2015). *The digital humanities: A primer for students and scholars*. Cambridge University Press

Paul Guyer (2015). Kritik, Kant-Lexikon(Online), Berlin/New York: Walter de Gruyter, 1303쪽

Glymour, C., Ford, K., & Hayes, P. (2000). “The prehistory of android epistemology”. *Artificial Intelligence: Critical Concepts, 1*, 113쪽

Haraway, D. (1995). *in Wolfe*, 36. Quoted from Posthumanism, 190쪽

Heidegger, M. (2000). “Frage nach der Technik”, *Gesamtausgabe Bd. 7,* Vittorio Klostermann, Frankfurt am Main, 8쪽

Horkheimer, M. (1988). Psychologie und Soziologie im Werk Wilhelm Diltheys, Gesammelte Schriften, Bd. IV, Fischer, 356쪽

Horkheimer, M. (1991). Zur Kritik der instrumentellen Vernunft, GS Bd. VI, Fischer, 27쪽

Kant, I. (1900 ff.) Kritik der reinen Vernunft, in: *Kants gesammelte Schriften(Sog. Akademie-Ausgabe)*, Walter de Gruyter

Kieras, D & Holyoak, K (1987). *Encyclopedia of Artificial Intelligence: Artificial Intelligence*. John Wileys& Sons

Kurzweil, R. (1999). “The Coming Merging of Mind and Machine-The accelerating pace of technological progress means that our intelligent creations will soon eclipse us--and that their creations will eventually eclipse”. *Scientific American*, (2), 56-61쪽

Nekrašas, E. (2001). “Pragmatism and positivism”. *Problemos*, 59, 41-52쪽

Orth, E. W. (1984). “Einleitung: Dilthey und der Wandel des Philosophiebegriffs seit dem 19. Jahrhundert”. *Phänomenologische Forschungen, 16*, 7-23쪽

Rich, E. (1987). *Encyclopedia of Artificial Intelligence: Artificial Intelligence.* John Wileys& Sons

Sorgner, S. L. (2009). “Nietzsche, the overhuman, and transhumanism”. *Journal of Evolution and Technology, 20*(1), 29-42쪽

A Sketch of Critique of Artificial Reason

Hyeongjoo Kim (Chung-Ang Univ.)

This study purports a critique of artificial reason toward the age of artificial intelligence and examines its academic possibility. To this end, I will outline the genealogy of the project of a critique of reason by reviewing the projects of critique of reason performed by I. Kant, W. Dilthey, and M. Horkheimer in turn. Through this, it will be clarified how Kant's concept of a critique of reason is revealed in different eras. In the process, it will be also argued that the meaning of the critique of reason includes, as a self-critique of reason, the analysis of the elements constituting reason and, as a critique of reason for others, the critique of scholarship in this era and the zeitgeist, based on an insight that the fundamental concept of the genealogy of critique of reason is identified with the restoration of autonomous reason and the pursuit of the academic system. All this will be applied to the fields of AI humanities. Specifically, according to the first role of critique, AI humanities are categorized into AI philosophy, post-humanism, digital humanities, and AI ethics. And if these categories of AI humanities are criticized from the point of view of "the critique of reason", such as "the pursuit of a system" and "the restoration of autonomous reason," it is argued that empiricist quantificationism is derived as a common feature of the four disciplines. Finally, it will be argued that it is not desirable to apply this position in a one-size-fits-all manner to all academic fields including the humanities. As an alternative, a critique of artificial

reason proposes the tautological anthropocentrism, which evaluates tautological anthropocentrism as a new canon that AI ethics should aim for.

Key words: Artificial Intelligence, Artificial Intelligence Ethics, Critique of Artificial Reason, Kant, Dilthey, Horkheimer

김형주 e-mail: godwithhj@gmail.com

투 고 일	2022년 01월 29일
심 사 일	2022년 02월 21일
게재확정	2022년 02월 22일