『人文科學』 제129집 2023년 12월
DOI http://doi.org/10.23017/inmun.2023.129..93

수학적 추상화가 유발하는 인공지능 오작동
- 윤리적 감성과 감정에 대한 레비나스의 고찰에 입각한 기술적 대안 탐구*

박욱주**

I. 들어가는 말

디지털 문화의 사회적 영향을 헤아리는 데 주력해온 연구자 루크 먼(Luke Munn)은 2022년 "인공지능 윤리의 무용성(The Uselessness of AI Ethics)"이라는 논문에서 현재까지 인문학 분야에서 수행된 인공지능 윤리 담론의 중대한 한계를 지적했다. 그는 작금의 인공지능 윤리 담론이 세 가지의 난관에 처해 있다고 진단한다. 첫째, 인공지능 윤리가 제시하는 목표가 전반적으로 모호하며, 둘째, 인공지능 산업계의 이해관계를 도외시하는 성향이 있고, 셋째, 현재 발전을 거듭하고 있는 실제 기술에 적용할 만한 기술적 전문성이 결여되

* 이 논문은 2019년 대한민국 교육부와 한국연구재단의 인문사회분야 신진연구자지원사업의 지원을 받아 수행된 연구임(NRF-2019S1A5A8033570)
** 연세대학교 연합신학대학원 연구교수

어 있다는 것이다(Munn, 2022, 869-870). 기술적으로 유효한 윤리적 실천원리를 제시하는 일은 모든 과학기술윤리 담론의 최종 목표라고 말할 수 있다. 먼이 지적한 것처럼, 인공지능 윤리 담론의 결실 역시 실제 기술적 적용이 가능한 통찰을 포함해야 한다. 그런데 인공지능 윤리는 여러 다른 과학기술윤리의 지류와 비교했을 때 한 가지 두드러지는 고유한 특징이 있다. 바로 인공 에이전트(artificial agent)[1]의 자율성(autonomy)이 초래하는 오작동에 대한 고찰의 필요성이다. 인간에 의해 인공 에이전트에게 분여된 자율성이 그 본래의 존재적, 기능적 목적을 벗어나 독단성으로 변이되어 인간의 삶에 위해를 가하는 오작동 문제는 인간이 인공 에이전트에게 더 많은 결정권을 허용할수록 그 심각성이 가중될 수밖에 없다.

본고에서 지적하는 인공지능 오작동의 근본 원인은 수학적 추상화(mathematical abstraction)이다. 인공지능 컴퓨팅은 현상들 사이의 상관관계를 수리논리 혹은 통계 형태의 표상으로 변환하는 방식으로 이루어진다. 이 방식은 필연적으로 현상적 실재의 다층성을 상당부분 배제하는 환원적 인지로 귀결된다. 이 환원의 과정에서 외면된 현상들의 개별적 본질들 가운데는 인지주체가 현실의 복잡한 문제상황에 적절하게 대처하는 데 필수적인 이해의 근거가 자리잡고 있는 경우가 다반사이다. 결국 수학적 추상화로 인한 환원적 인지는 인공 에이전트가 그 목적에 합당한 기능을 수행하는 데 어려움을 주는 근본 원인이 된다. 인공지능 기술의 혁신은 이 난관에 얼마나 적절하게 대응

1) '에이전트'란 독자적인 의지와 자율적 결정권을 가진 지적 행위주체를 말하는 것으로, 통상적으로는 확고한 자기의식을 가진 인간 개개인이나 특정한 목적을 위해 인간들이 모여 구성한 집단을 지시한다. 자기의식이 존재하지 않는, 그리고 인간 수준의 자유의지를 갖지 못하는 인공지능 시스템을 하나의 온전한 에이전트로 규정하는 데는 분명 논란의 여지가 있다. 그러나 인공지능 시스템이 논리적 추론과 통계적 연산을 통해 일정범위 안에서 제한적으로나마 실제적인(actual) 자율성을 행사하여 환경에 변화를 주고 적응하는 이상, 잠정적으로 하나의 인공지능 시스템을 에이전트로 규정하는 것이 전적으로 불가능하지는 않다. 즉 '지능'이나 '마음'을 어느 범위까지 정의하느냐에 따라 에이전트의 규정 범위 역시 크게 달라진다(Schmidt, 2007, 357). 이런 이유로 인공 에이전트라는 용어는 주로 컴퓨터공학 분야에서 큰 거부감 없이 사용되고 있다(Picard et al., 2023, 217).

하느냐에 달려 있다고 봐도 과언이 아니다.

본고는 인공지능 컴퓨팅 방식의 핵심인 수학적 추상화가 다양한 인공지능 오작동을 일으키는 근본 원인이라는 사실을 두 차례의 실제 사례를 통해 살펴보며, 이런 오작동이 궁극적으로 현상적 실재를 가소적으로 구성함에 따라 유발된다는 것을 논증하려 한다. 그리고 인공 에이전트에 의해 수행되는 추상화된 인지가 타인과 마주하는 감성적 현실을 외면하게 만들어 인공지능이 윤리적 책임의 현상적 발원지에서 멀어지게 만드는 주된 원인이 된다는 사실을 조명한다. 이를 위해 에마뉘엘 레비나스(Emmanuel Levinas)의 타자윤리를 관통하는 대상화 및 추상화에 대한 비판적 통찰을 주된 반성의 준거로 채택한다. 이어서 최근 각광받고 있는 감정 인공지능(Emotional AI) 기술의 특징을 살펴보고, 이 기술이 현재 컴퓨터 엔지니어들과 인공지능 서비스 사용자들이 직면하고 있는 오작동 및 윤리적 결함 문제에 보다 적절하게 대응하도록 돕는 하나의 대안이 될 가능성을 살펴본다.

II. 수학적 추상화의 한계: 현상적 실재와 인지된 대상의 불가피한 괴리

먼은 인공지능 개발과 관련해서 현재까지 도출된 다양한 윤리적 원리들이 기술개발 현실에 실제 적용할 수 없는 비전문성의 문제를 내포하고 있다고 지적하면서 이런 원리들을 “실효성 없는 원리들(toothless principles)”이라고 명명한다(Munn, 2022, 871). 이를 통해 그는 인공지능 윤리 담론이 여러 규범적인 이념들을 제시하고 있기는 하나 이 이념들에 담긴 가치와 원리들을 기술적으로 구현할 만한 컴퓨터공학 체계나 절차를 정립하는 데 이르지 못하고 있는 현 상황을 보여주려 한다. 그런데 먼이 지적하는 이 인공지능 윤리 담론의 한계는 역으로 현재 심층학습 중심으로 발전되고 있는 인공지능 기술 자체의 한계를 보여주기도 한다. 현재의 컴퓨터 공학 수준으로 개발되는 인공 에이전트 대다수가 기능과 윤리 양면으로 아직 충분히 신뢰할만한 인지력 및

판단력을 갖추지 못하고 있고, 그래서 인공지능 윤리 담론이 여기에 실제적으로 공헌할 길을 찾아야 할 과제가 남아있는 것이다.

먼의 논의를 유념해볼 때, 일단 인공지능 윤리 담론의 결실을 실제 기술개발 현실에 적용하는 주된 목표 가운데 하나로 인공지능 오작동 방지를 지목할 수 있다. 인공 에이전트가 기능적으로든 윤리적으로든 본래 지정된 목표에 미달하는, 혹은 목표를 거스르는 결과값을 내놓는 상황은 안전하고 효율적인 인간-인공지능 상호작용을 가장 크게 저해하는 요인 중 하나이다. 사산카 찬다(Sasanka S. Chanda)와 데바락 바네르지(Debarag N. Banerjee)에 의하면, 기능적 관점으로 볼 때 인공지능 오작동은 크게 두 부류로 나뉜다. 첫째, 이행해야 할 바를 이행하지 않는 "누락 오작동(ommisision error)," 둘째, 이행하지 말아야 할 것을 이행하는 "과실 오작동(commission error)"이다(Chanda & Banerjee, 2022, 1). 그리고 이 두 부류의 오작동이 발생하는 데는 다시금 크게 두 가지의 구조적 원인이 존재한다. 첫째, 문제 상황에 대응하기에 적절치 않은(혹은 충분치 않은) 입력값 데이터로 연산과 인지를 수행하는 것, 둘째, 입력값이 적절하게 주어져도 신경망 알고리즘 자체에 결함이 있어 문제 상황에 적절히 대응하지 못하는 것이다(Chanda & Banerjee, 2022, 2).

다음 오작동 사례는 입력값 혹은 신경망 결함이 어떤 식으로 누락 오작동을 초래하는지 보여준다. 2017년 11월 8일, 미국 라스베이거스 시 당국은 비교적 단순한 구간을 반복 운행하는 자율주행 셔틀을 도입해 운영을 개시했다. 하지만 이 셔틀은 운행 첫날 단 두 시간 만에 저속 추돌사고를 일으켰다. 이 자율주행 차량은 한 명의 운전요원과 여덟 명의 승객을 태우고 있었으며, 운전요원은 필요할 때 셔틀 컨트롤러를 통해 운행에 관여할 수 있도록 대기 중이었다. 이 차가 스트립 지역의 한 교차로를 지나갈 무렵 측면에서 한 대의 택배 트럭이 같은 교차로로 진입하고 있었다. 자율주행 셔틀은 트럭을 인지하지 못한 채 트럭이 주행하는 경로로 밀고 들어갔고, 이 상황을 알아챈 셔틀 운전요원은 크게 놀라 운행을 중단하려 했으나 컨트롤러가 운전요원의 손에 곧장 닿는 자리에 설치되어 있지 않았다. 트럭이 셔틀 앞쪽 범퍼 측면에 추돌

하자 셔틀은 비로소 운행을 멈췄다. 저속 추돌이라 큰 피해는 없었으나 운전요원과 탑승자들은 순간적으로 큰 심적 충격을 받았다. 개발사는 상대방 트럭 운전자의 과실을 주장했으나, 미 연방 교통안전 위원회(National Transportation Safety Board, NTSB)는 1년이 넘는 조사 끝에 이 사고가 자율주행 시스템의 실책이라는 결론을 내렸다. 해당 시스템이 측면에서 다가오는 트럭과 충분한 안전거리를 유지하지 않은 점, 그리고 컨트롤러의 위치가 운전자와 멀어 위급할 때 운전요원이 컨트롤러를 즉시 사용할 수 없었던 점이 이 사고의 주원인으로 지목되었다(National Transportation Safety Board, 2019, 15-16).

찬다와 바네르지는 이 추돌 사건을 적절한 입력값이 충분하게 주어지지 않아 발생한 누락 오작동으로 분류한다. 해당 인공 에이전트가 이 사고에 해당하는 시나리오 유형을 학습할 수 있도록 유사한 도로 상황과 사고 사례에 대해 충분한 학습을 수행했어야 하는데 개발사 측에서 이 점을 간파하지 못했다. 특정 각도에서 특정한 형태의 차량이 접근할 때 어느 정도의 안전거리를 확보해야 하는지 연산하는 시나리오가 알고리즘에 포함되어 있지 않았던 것이다. 그래서 이 자율주행 시스템은 당시의 상황이 위험하다는 것을 인지하지 못해 누락 오작동을 일으켰다(Chanda & Banerjee, 2022, 3).

이런 자율주행 시스템의 누락 오작동은 디지털 컴퓨팅에 바탕을 둔 인공지능 본연의 한계에 기인한 것이다. 이 한계는 주로 수학적 추상화 방식에 대한 고찰을 통해 보다 심층적으로 이해할 수 있다. 로렌자 사이타(Lorenza Saitta)와 진-다니엘 주커(Jean-Daniel Zucker)가 정리한 바에 따르면 추상화란 기본적으로 배제 및 환원의 의미를 가진 행위로 정의할 수 있다. 추상화(abstraction)라는 용어는 '제거하다' 혹은 '추출하다'라는 의미를 지닌 그리스어 '아파이레시스(ἀφαίρεσις)'로부터 유래되었다. 이로부터 파생된 여러 의미 가운데 가장 주목해야 할 추상화의 두 가지 특징은 "사태들(events)"의 배제와 "관념적 본질(the essentials)"의 추출일 것이다(Saitta & Zucker, 2013, 1). 추상화 방식은 철학, 수학, 예술 분야에서 주로 채택되어 왔는데, 인공지능 설계와 관련해서는 특히 수학적 추상화 방식을 눈여겨볼 필요가 있다. 수학은 추상적

인 학문으로서 “감각 세계로부터 거리가 먼 것”(Saitta & Zucker, 2013, 2)을 주로 다룬다. 수학은 기본적으로 현실에 대한 표상을 조성하는 학문 분과로서 인간의 직접적인 현상체험과 한 단계 떨어져 있는 규칙과 논리를 모색하는 데 주력한다. 수학 내부에 직관이라는 개념이 없는 것은 아니지만 수학적 직관은 오늘날 현상학적 관점으로 보는 직관[2]과 그 의미가 분명하게 구별된다. 수학적 직관은 칸트적 의미에서의 직관, 즉 인간의 감관 속에 선험적으로 자리잡고 있는 현상 감지의 수학적 형식을 의미한다(이종권, 2005, 192). 그래서 수학의 직관주의 또한 감각적으로 체험하는 날것의 현상 그 자체보다는 그 현상을 감지하는 인간 정신 내면의 수학적 원리와 그것에 의해 구성되고 종합된 표상들을 직접적인 연구 대상으로 지목한다(Heyting, 1983, 53).

디지털 컴퓨팅은 현상을 수학적 대상, 좀 더 구체적으로 말하자면 “디지털 대상(digital object)”으로 전환하는 것으로부터 시작된다(Hui, 2016, 48). 디지털 컴퓨터가 처리하는 모든 현상은 궁극적으로 0과 1 두 개의 숫자들(digits)로 변환된 것이다. 인공지능 컴퓨팅은 궁극적으로 이 디지털 컴퓨팅을 바탕으로 일정한 의미 창출의 자율성이 부여된 “의미의 망(semantic web)” 속 구조에 맞게 라벨링된 입력값을 수용한 다음 수리적 혹은 통계적 연산을 수행한다(Hui, 2016, 50). 다시 말해 인공 에이전트가 인지하는 모든 현상은 디지털 방식으로 표상화된 것들이며, 이 표상들에서 유의미한 정보를 추출해내는 알고리즘 또한 그 안에서 일정 기준 이상의 상관관계를 가진 것들만을 추려내는

2) 현대 현상학의 관점에서 직관이란 과거 칸트가 제시한 것과는 크게 다른 것으로 규명된다. 현대 현상학 정착에 크게 기여한 에드문트 후설(Edmund Husserl)과 마르틴 하이데거(Martin Heidegger)의 직관에 대한 성격규정 방식을 보면 오늘날 갱신된 직관 개념의 특성에 대해 분명하게 알 수 있다. 후설은 현상에 대한 인간의 직관 안에 각 개인의 고유한 지향성이 반영되어 있으므로 모든 직관은 단순 객관화하기 어려운 개별적 본질을 내포하는 체험으로 규명된다고 설명한다(Husserl, 2009, 61-62). 하이데거는 직관이 어떤 고정적인 선험적 형식에 따라 일어나는 것이 아니라 각 현존재의 삶의 정황에 적실하게 유동적으로 체감되는 시공간적 세계에 대한 고유하고 자유로운 체험이라고 밝히고 있다(Heidegger, 2003a, 230-231). 두 사람의 견해를 종합해본다면 현대 현상학에서 현상에 대한 직관이란 개별성, 체험적 직접성, 그리고 형식과 규칙에 구애받지 않는 감성적 유동성(혹은 자유)을 그 대표적인 특징으로 삼는 것으로 규명된다.

추상화를 수행하도록 설계되어 있다. 인공지능은 이런 추상화를 통해 문제상황에 대처하기 위한 인지의 단순성과 효율성을 확보한다(Saitta & Zucker, 2013, 6-7). 잡다하고 처리하기 어려운 형태의 현상들을 일정한 패턴으로 단순화하고, 이 단순화된 인지를 바탕으로 가장 효율적인 대응책을 수립해 문제해결의 효율성을 갖추는 것이다.

루치아노 플로리디(Luciano Floridi)는 컴퓨터가 현상을 인지하는 방식을 추상화 단계(levels of abstraction) 기법을 통해 상술한다. 플로리디에 따르면(2011, 48) 디지털 컴퓨팅은 변수(variable)와 가측치(observable)의 두 개념으로 현상을 연산 가능한 데이터로 전환한다. 변수는 어떤 현상에서 문제 해결에 필요한 것으로 상정된 측면을 수학적으로 다룰 수 있는 개념으로 기호화한 것이고, 가측치는 이 기호화된 현상의 일면에 수치로 가치를 부여하는 것을 말한다. 변수와 가측치를 통해 디지털 대상으로 전환된 현상은 이내 수리적으로 설정된 관계(relation)에 따라 보다 유의미한 정보(information)로 탈바꿈된다(Floridi, 2011, 53). 변수 x에 배정된 가측치들의 집합을 A, 변수 y에 배정된 가측치들의 집합을 B라고 정하면, 두 변수에 속한 가측치들 사이의 특정한 관계 사례들을 모아놓은 집합 R은 데카르트 곱(Cartesian product)인 A × B의 부분 집합이 된다. 이는 다음의 수식으로 표현될 수 있다(Saitta & Zucker, 2013, 72).

$$COV(R) = \{ (x, y) \mid x \in A, y \in B \text{ and } R(x, y) \}$$

이 수식은 집합 A와 B에 속한 확률변수 x와 y 사이의 공분산(covariance) 식으로서, 변수 x와 y에 배정된 모든 가측치들 사이의 특정한 상관관계 R의 정도를 나타낸다. 이 식의 결과값이 클수록 x와 y 사이의 상관관계는 높아진다. 조건에 맞는 특정 가측치들이 결합되면 현실에서 보다 유의미한 가치를 갖는다는 뜻이다. 추상화 단계 기법은 이처럼 특정한 요구사항에 따라 현상적으로 가장 높은 가치를 갖는 최적의 상태를 가측치 사이의 상관관계 정도를

구하는 수식으로 치환한다.

플로리디는 방금 소개한 변수, 가측치, 그리고 관계까지의 추상화 단계들이 어떻게 실제 디지털 컴퓨팅의 추상화에 적용되는지 쉽게 이해할 수 있도록 포도주의 가치를 판단하는 사례를 설명한다. 개인적인 기호의 관점에서 '좋은 맛을 내는 포도주'를 선택하려 할 경우, 디지털 컴퓨팅으로 이 판단을 추상화하기 위해서는 포도주를 선택하는 데 중요하게 여겨지는 유형화된 변수들을 설정하고 거기에 가측치를 부여해야 한다. 가령 "향취, 빛깔, 색채, 산도, 포도맛의 정도, 뒷맛의 강도" 등을 유형화된 변수로 설정할 수 있을 것이다(Floridi, 2011, 52). 다음으로 각 변수의 가측치들 가운데 최적의 조합 사례를 선정하기 위해 특정한 연산법칙에 따라 가측치들 간의 관계 집합 R의 원소 전체를(혹은 특정한 필터링 기법에 의해 일부를) 검증한다. 이와 달리, 만일 경제적인 판단의 관점에서 '구매에 유리한 포도주'를 선택하려 한다면 방금 전과는 다른 유형의 변수들이 설정된다. 가령 "생산자, 생산지, 숙성연도, 공급자, 와인의 양, 가격" 등이 최적의 포도주 선정을 위한 변수로 선정될 수 있다(Floridi, 2011, 52). 이 경우에도 마찬가지로 각 변수에 배정된 가측치들의 관계 조합을 검증하는 연산을 통해 최적의 가측치 조합을 선정한다.

플로리디가 설명하고 예시한 추상화 단계 기법에서 주목해야 할 사안은 추상화가 진행될수록 인지 주체가 각 단계별로 직관적이고 아날로그적인 현상체험에서 점점 더 멀어진다는 사실이다. 유형화된 변수를 선정하는 단계에서는 각 변수가 대표하는 현상의 일면들만 인지에 고려되고 나머지는 배제된다. 가측치를 부여하는 단계에서는 감성적 직관이 수치로 변환되며, 관계 검증 단계에서는 현상체험의 여러 요소들 사이에서 확인되는 연관성이 패턴화되고 거기에 다시금 수치로 확인할 수 있는 가치가 부여된다. 이처럼 인간이 오감으로 체험하는 현상의 복잡다단하고 종합적인 측면들을 배제와 패턴화를 통해 수리적 표상으로 환원하는 행위는 다분히 데카르트적인 인식의 원리를 따른 것이다. 르네 데카르트(René Descartes)는 존재를 우선 연장된 실체(*res extensa*)로 규정한 다음 기존 서구 철학에서 규정한 여러 존재의 범주들 전체

를 기하학적 연장의 형태로 환원해서 표상화할 것을 권고한다(Descartes, 2013, 111). 여기서 연장의 형태로 환원되는 범주들 가운데는 당연하게도 관계의 범주가 포함되어 있다. 그리고 다음으로는 존재의 범주들 전체를 기하학적으로 표상화하는 행위의 정당성 확보를 위해 회의하는 순간 절대적으로 부정할 수 없이 존재하는 주체의 사고하는 실체(*res cogitans*), 코기토(*cogito*)의 존재적 확실성을 역설한다(Descartes, 2013, 185). 현상을 통해 감지된 존재는 엄밀한 수학적 표상화에 의거해서 주체의 사고 속에, 혹은 주체의 사고에 의존해서 그 실재성을 확보한다. 따라서 데카르트의 관점으로 본다면 디지털 컴퓨팅을 위한 추상화 단계 기법은 존재의 실재를 가장 엄밀하게 인식하는 모범적 사고의 한 방편이 될 가능성을 충분하게 담지하고 있다.

이러한 사고는 인식 주체가 특정한 형식을 통해 자의적으로 주위세계에 대한 표상을 지어내고 거기에 진리 인식의 기반을 둔다는 점에서 맥락상 임마누엘 칸트(Immanuel Kant)가 설명한 직관형식과 상당부분 일치하는 면이 있다. 물론 칸트는 데카르트처럼 존재의 실재 혹은 물자체를 인식하는 것이 가능하다고까지 주장하지는 않지만, 인간이 접근할 수 있는 최대치의 진리란 궁극적으로 인간이 현상을 받아들이는 감성[3]과 지성의 형식에 최종적인 근거를 두고 있다고 믿는다(Kant, 2006, 239-240). 이러한 점에서는 칸트 역시 데카르트와 유사하게 인간 인식을 주체 내부로 내재화하고 있다. 앞서 언급했던 것처럼 수학적 직관은 이 칸트적 직관 개념을 계승하고 있다. 수학적 직관주의 또한 큰 틀에서 본다면 감각적 현상체험 그 자체보다는 이 현상체험을 계기삼아 인간이 자기의 정신 내면에 그려낸 표상을 수학적 진리 탐구의 기초근거

3) 감성과 감정은 일상에서 자주 혼용되는 용어이지만 인식론과 현상학에서는 양자의 의미가 대개 구별되며 본 논의에서도 이런 구별을 따른다. 감성은 감관으로 느끼는 체험, 직관을 가능하게 하는 능력을 뜻하는 반면(Kant, 2006, 239), 감정은 이 직관이 대상화되기 전, 지성의 판단을 거치기 전 느껴지는 기분을 말한다. 감정으로서의 기분은 실존론적 현상학에서는 인간이 자기 존재 방식의 적실성을 막연하지만 진득하게 판별하는 능력으로도 규명된다(Heidegger, 2001, 112). 양자의 관계를 따지자면 감성은 감정을 유발하는 원인이 되고, 감정은 감성의 산물이라고 볼 수 있다.

로 상정하고 있다.

현상에서 여러 단계 멀어진 수학적 추상화에 대한 이런 인식론적 고찰을 위에 언급한 자율주행 셔틀의 오작동 사례에 적용해볼 때, 가장 먼저 주목해야 할 사실은 찬다와 바네르지 두 연구자가 이 오작동 사례의 원인으로 불충분한 입력값을 지목했다는 점이다. 이들은 해당 자율주행 시스템 신경망의 신뢰도에 대해서는 별다른 의구심을 갖지 않는다. 플로리디의 추상화 단계 기법으로 풀이하자면, 적어도 이 인공지능 자율주행 시스템은 안전하고 적법한 도로 주행에 필요한 여러 변수와 그 가측치들 사이의 관계 집합 R을 규정하는 연산 공식의 구조적 체계에 별다른 하자가 없다고 보는 것이다. 즉 찬다와 바네르지는 해당 사고를 유발한 변수들이 비슷하게 부각되는 상황들에 관한 데이터가 입력값으로 충분하게 제공된다면 이 자율주행 시스템은 더 이상 같은 유형의 사고를 내지 않을 것이라고 전망한다(Chanda & Banerjee, 2022, 3).

찬다와 바네르지의 기술적 전망과 관련해서 주목해야 할 점은 인간 운전자에는 굳이 필요하지 않은 면밀하고 수고스러운 학습의 과정이 인공 에이전트에게는 필수적으로 요구된다는 점이다. 셔틀에 탑승하고 있던 운전요원은 사고가 날 것이라는 것을 충분하게 예견하고 크게 놀라 즉시 운전에 관여하려 했다. 즉, 해당 사고 상황은 인간 운전자가 안전운전에 필요한 능력을 어렵지 않게 발휘했을 만한 상황이었던 것이다. 이 사실이 시사하는 바는 인공 에이전트가 디지털 대상으로 변환된 현상을 입력값으로 수용할 때 일정한 측면에서 그 입력값에 담긴 의미를 적절하게 포착하지 못한다는 점이다. 즉, 적절한 변수의 설정, 충분한 가측치의 제공, 그리고 목표를 적중하는 가측치들 간의 연관성 검토, 이 세 가지 가운데 하나라도 어긋나는 요소가 있다면 원래 목표로 삼은 문제해결 능력의 발휘는 불가능하며 수시로 오작동을 낼 수밖에 없다. 만약 해당 에이전트가 처리하는 문제상황이 확실하게 통제할 수 있는 변수들로만 구성되어 있다면 인공지능 오작동은 발생하지 않을 것이다. 사이버 공간 안에서 규칙이 명확하게 정해진 문제들을 해결할 때 인공지능은 대단한 성능을 발휘한다. 알파고나 최근 두각을 드러내고 있는 이미지 생성 AI 등이

대표적인 사례라고 볼 수 있다. 돌발변수가 전적으로 제거된 비교적 단순한 규칙들로 구성된 가상공간 안에서 기존에 학습된 내용을 바탕으로 충분히 처리할 수 있는 문제상황들만 상대하는 인공지능의 성능은 동일한 사안에 대한 인간의 문제해결 능력을 압도한다. 그러나 통제되거나 예상되는 범위 바깥의 우연하고 돌발적인 현상들이 잡다하게 일어나는 현실의 상황, 여기에 더해 인간 개개인의 자유로운 행동들까지 관여되는 물리적인 인간-인공지능 상호작용의 상황에서 인공 에이전트의 성능은 인간에 비해 여러 모로 부족함을 드러낸다(Washizaki, 2022, 45). 이는 인간이 감성과 지성 모두를 활용해 직관적 현상체험에 대응하는 것과 달리 인공 에이전트는 수학적 추상화에 의거해서 오로지 지적인 인지, 그것도 지성의 제한된 영역만 활용하는 인지를 수행하고 이를 바탕으로 문제상황에 대응하는 결과값을 내놓기 때문이다.

수학적 추상화에 의존하는 인공지능 컴퓨팅의 이러한 한계는 기본적으로 누락 오작동이 아닌 과실 오작동에서도 동일하게 확인된다. 찬다와 바네르지가 예시한 것처럼, 2018년 10월의 라이온 항공 610편 추락사고와 2019년 3월의 에티오피아 항공 302편 추락사고는 모두 보잉 사가 개발한 인공지능 시스템 MCAS(Maneuvering Characteristics Augmentation System)가 문제가 되어 발생했다(Chanda & Banerjee, 2022, 6). 이 시스템은 보잉 사가 새로 개발한 보잉 737 맥스 기체의 항공역학적 약점을 보완하기 위해 개발되었다. 이 신형 항공기는 저속비행시 기수가 위로 들리며 실속이 일어나는 구조적인 문제를 안고 있었는데, 보잉 사는 기체 구조를 새로 설계하는 데 드는 비용을 아끼기 위해 새로운 항공기 소프트웨어 MCAS를 개발해 설치했다. MCAS는 수동비행 상태에서 항공기 속도가 일정 수준 아래로 떨어지면 조종사의 의향과 상관없이 자율적으로 작동하는 제한적 인공지능 시스템이었다(Henderson, Harbour, & Cohen, 2022, 1). 이 시스템은 다중 은닉층을 포함하는 인공신경망 기반의 심층학습 인공지능은 아니다(Schildt, 2020, 143). 하지만 명제적 언어논리를 수리적으로 구현하여 일정한 범위 내에서 자율적으로 전문가의 소견과 같은 결과값을 내놓는다는 점에서 인공지능의 일종인 전문가 시스템에

속한다. 이 항공기용 인공지능 시스템은 AOA센서(받음각 센서, 즉 기체의 날개와 기류가 이루는 각도를 측정하는 센서)가 정상 작동하는 상황에서는 기수가 들리는 것을 방지하고 기체의 실속을 막는 유용한 조종 보조 역할을 담당한다. 그러나 AOA센서가 오작동하는 경우, 잘못 입력된 받음각 수치에 따라 기체가 정상적으로 운행하는 상황에서 기수를 낮춰 추락 위험에 빠지게 한다(Mongan & Kohli, 2020, 1). 라이온 항공 610편과 에티오피아 항공 302편 추락사고 모두 AOA센서가 인공지능 시스템인 MCAS에게 부정확한 입력값을 제공해서 발생한 사고인 것으로 밝혀졌다(Naor et al., 2020, 1, 4). 찬다와 바네르지가 지적한 대로, 굳이 내놓지 말아야 할 결과값을 내놓고 시스템의 독단성에 의거해서 그것을 일방적으로 수행한 과실 오작동이었던 것이다.

표면적으로 이 사고는 인공지능 시스템이 물리적 현상을 가장 직접적으로 마주하는 센서에서 오작동이 발생했고 이로 인해 생성된 부적절한 입력값을 제공받은 알고리즘이 잘못된 결과값을 도출한 탓에 발생했다. 그러나 홍준 장(Hongjun Zhang)과 바이챠오 황(Baiqiao Huang)이 지적한 것처럼(2020, 709), 이것은 단순히 센서 오작동의 문제가 아니라, 센서 오작동에 의해 주어진 부정확한 입력값을 맹목적으로 인지의 근간으로 여긴 시스템 전체의 문제로 취급되어야 한다. 장과 황은 이 시스템이 다음 세 가지의 결함을 가지고 있다고 분석한다. 첫째, 현상 인지 방식이 "과도하게 단순"(Zhang & Huang, 2020, 709)하다. 하나의 센서 고장이 전체 인지를 좌우할 정도로 현상과 접점을 구성하는 방식이 심히 단편적이라는 것이다. 둘째, 인공 에이전트와 "인간의 협력이 결여"(Zhang & Huang, 2020, 709)되어 있다. 프로그램이 인간 조종사의 피드백을 무시하는 방식으로 작동하도록 설계되어 있다는 것이다. 셋째, 현실 인지에 오류가 발생했을 때 "오류를 수정하는 메카니즘이 부재"(Zhang & Huang, 2020, 709)하다. 인간 조종사의 피드백을 받지 않는다면 자체적으로라도 가능한 결과값 오류를 수정하도록 보완 역할을 담당하는 알고리즘 코드와 이를 실행하는 하드웨어 설비가 갖추어져 있어야 하는데 그렇지 못하다는 것이다.

장과 황 두 연구자가 지적한 세 가지 결함들은 수학적 추상화의 두 가지 한계, 즉 현상에서의 괴리와 그로 인한 독단적 인지가 현실과 부조화를 일으킬 때 나타나는 전형적인 양상이라고 볼 수 있다. 추상화 단계의 틀을 가지고 살펴보면, 보잉 사의 MCAS 오작동 사례는 우선 변수 설정 방식부터 문제점을 드러냈다. 받음각 측정을 위한 센서가 단 하나였고 이 센서의 오작동을 보완할 만한 다른 보완책이 없었다. 보잉 737 맥스의 기수에는 두 개의 받음각 측정 센서가 있지만, MCAS의 발동을 위해서는 이 중 하나의 센서만 사용되도록 설계되어 있었다(Zhang & Huang, 2020, 709). 두 센서 중 하나가 오작동을 일으켜 서로 다른 각도를 측정했을 때 이를 조종사에게 알리는 장치가 개발되어 있었으나 보잉 사는 이 장치를 추가적인 비용을 낼 때만 설치하는 업그레이드 옵션으로 지정했기 때문에 대다수의 737 맥스 기체에는 이 장치가 설치되지 않았다. 게다가 보잉 사는 MCAS와 관련해서 이 장치가 중요하다는 사실을 항공사와 조종사들에게 충분히 고지하거나 학습시키지도 않았다(Rose, 2022, 113). 이로 인해 이 기체를 조종하는 조종사 대부분이 MCAS의 존재를 몰랐고 MCAS가 잘못된 상황에서 기수를 내릴 때 이를 수동으로 무효화(override)하는 방법 또한 알지 못했다. 비행안전에 직결되는 중요한 판단을 내리는 데 단 하나의 변수만 입력값으로 수용하도록 설계되어 있었고, 결과값에 대해 인간 조종사가 적절히 피드백을 내리고 해당 시스템을 통제할 수 있게 하는 교육 또한 전혀 이루어지지 않았던 것이다. 다시 말해서 MCAS는 하나의 변수에 배정되는 가측치가 부정확하면 시스템 전체가 오작동할 수 밖에 없는 단순하고 부실한 관계 검증 구조를 갖고 있었다. 따라서 이 사고는 단지 센서 결함의 문제로만 발생한 것이 아니라 해당 인공지능 시스템 전체가 구조적으로 안전을 도외시한 탓에 발생한 것이라고 보는 것이 타당하다.

지금까지 살펴보았듯이, 라스 베이거스 자율주행 셔틀의 누락 오작동, 그리고 보잉 737 맥스 MCAS의 과실 오작동은 모두 부적절한 변수, 가측치, 그리고 관계 검증 구조 때문에 수학적 추상화가 현상적 실재와 부조화를 일으킨 데서 발생한 것으로, 적절한 수학적 추상화의 기술적 난이도가 대단히 높다는

것을 보여준다. 특히 인공지능 시스템 개발자가 상정한 범위 바깥의 돌발적인 상황이 발생하면 아무리 인공 에이전트가 전문가 수준으로 고도화된 인지를 수행하거나 막대한 양의 학습을 수행한다 하더라도 문제상황에 적절히 대응하지 못한다는 것을 볼 수 있다. 물론 인공지능 기술은 정체되어 있지 않고 빠르게 발전 중이며, 상기 오작동 사례들은 학습의 강화나 시스템 보강을 통해 점차 극복되고 있다. 그러나 이런 보완책이 마련되기까지 해당 기술로 인해 여러 인간 사용자가 사고 위기를 겪거나 목숨을 잃는 일이 발생했다. 이를 기술발전 과정에 수반되는 단순한 실패 사례나 불운으로 취급하기에는 개발자나 서비스 제공자들의 무책임함이 심히 선명하게 부각된다. 인간 운전자나 조종사라면 당연하게 회피하거나 방지할 수 있었던 치명적인 사고를 인간에게 강요하는 수준의 시스템을 개발해 사용하게 했다는 데서 해당 인공지능 개발자들과 제작사의 도의적 책임은 간과할 수 없으리만치 명확하다. 따라서 이러한 류의 오작동 방지를 위해서는 면밀한 기술적 보완책에 더하여 심층적인 윤리적 반성도 함께 요구된다.

III. 윤리적 감정의 구현: 인공지능 컴퓨팅의 타자윤리적 반성과 기술적 대안

인공지능을 활용하는 것, 그것은 여타의 도구를 사용하는 것과는 사뭇 성격이 다른 일면을 가지고 있다. 인공 에이전트는 정해진 목적과 기능의 한계 안에서 수리적 기법에 따라 분여된 제한적 자율성을 가지고 있다. 이 자율성은 문제상황의 유동성에 적절하게 대응하도록 하는 필수 요소지만, 그로 인해 인공지능과 상호작용하는, 혹은 인공지능의 도움을 받는 인간의 일상적 자유를 심각하게 제한하기도 한다. 이런 문제는 특히 문제상황이 복잡해서 그에 대한 인공지능의 대응에 더 많은 자율성을 부여해야 할 때 주로 발생한다 (Russell, Moskowitz, & Raglin, 2017, 85). 상기 두 차례의 오작동 사례에서

확인되는 것처럼 인공지능의 자율성이 오작동을 낳고 그 오작동에 관여할 수 있는 인간의 자유로운 의지가 시스템의 구조적 결함에 의해 제한되거나 저지되는 순간 더 이상 바람직한 인간-인공지능 상호작용은 불가능해지고 이 상호작용 자체가 불안과 두려움, 그리고 실제적 고통을 유발하게 된다.

이 오작동 문제를 해결하기 위한 두 가지 접근 방향이 존재한다. 첫째는 인공지능의 성능 향상에 더욱 전념하는 것이다. 지금까지 발생한, 그리고 향후 발생할 오작동 사례 전체를 검토한 다음 수학적 추상화 기법을 더욱 정교하고 적절하게 보강해서 인공 에이전트가 개발자가 부여한 목표에서 벗어날 가능성, 최적의 결과값과 괴리된 오차값의 발생 가능성을 최소한도로 줄이는 것이다. 이는 심층학습이라는 컴퓨팅 기법이 기본적으로 지향하는 방향성이며, 현재 수많은 컴퓨터공학 연구자들과 기술기업들이 몰두하고 있는 방향성이기도 하다. 그러나 인공지능의 성능을 극단적으로 향상시켜 아예 비윤리적으로 작동할 가능성 자체를 최소화할 수 있다는 이 기술주의적(technocentric) 확신[4]은 현대 현상학의 관점과 타자윤리의 관점으로 볼 때 일말의 의구심을 불러일으킨다.

우선 인공 에이전트가 목표로 삼은 문제해결과 관련된 모든 상황, 특히 인간 개개인의 개별적 지향성이 관여된 인간-인공지능 상호작용의 다면적, 다층적 복잡함 속에서 과연 정교한 수학적 추상화만으로 문제상황에 온전히 대처할 수 있을지 의심스럽다. 인공지능의 성능에 대한 사용자의 요구가 보다 개인화되고 고도화될수록 해당 인공 에이전트가 대응해야 하는 문제상황은 더 유동적인 것이 되고, 그에 대응하기 위해 요구되는 자율성의 정도 또한 증대

4) 카람짓 길(Karamjit S. Gill)이 지적한 바에 의하면(2008, 218), 인공지능의 성능 강화를 통해 인간이 직면하는 대다수의 문제들을 이전보다 훨씬 더 효율적이고 바람직하게 해결할 수 있다는 기술주의적 확신의 이면에는 세계가 고정된 원인-결과 관계들에 지배되고 있다는 기술주의적 세계관이 자리잡고 있다. 이런 세계관에 따르면 세상의 모든 현상은 양적인 가측치로 치환이 가능하며, 이런 가측치들을 면밀하게 검토하면 특정한 문제 상황에 적합한 "유일한 최선의 해법(one best solution)"을 도출하는 것이 가능하다. 이처럼 유일한 해법을 추구하려는 단순화의 경향 때문에 기술주의적 확신에 의한 문제상황의 인지와 해법 도출 방식에는 불가피하게 현상 속 다양성이 축소되거나 환원되는 문제가 발생한다.

된다. 현실의 문제상황에서 인간은 상당히 복잡하면서도 비명시적인 인지와 판단의 능력을 발휘한다. 이 가운데는 명시적으로 의식되는 지적 판단보다 감성적인 직관에 의존해 순간적으로 발휘하는 대응력이 더 크게 작용하는 경우가 많다. 예를 들어 운전의 경우 인간은 안전한 주행에 필요한 능력 전부를 의식적으로 계산하지 않으면서도 눈으로 시야를 확보하고 귀로 음악을 들으며 손으로는 운전대를, 발로는 엑셀과 브레이크를 적절하게 조작한다. 이 다각적인 행위들은 특별한 저항이나 난관에 처하지 않는 이상 운전 중 대부분의 시간에 명시적으로 의식되지 않고 수행된다. 단지 인간은 운전을 하는 그 상황에 얕든 깊든 몰입한 가운데 필요한 업무를 수행한다. 이처럼 인간은 의식과 무의식을 모두 사용해서 자기를 둘러싼 문제상황을 인지하고 위험을 초래할 수 있는 잘못된 행위들, 즉 오류들을 배제하도록 스스로의 행동을 통제하는 "자율규제(self-regulation)"의 능력을 갖는다. 그리고 이 자율규제의 무의식적 측면은 문제상황에 직면해 있는 바로 그 순간 문제를 유발하는 오류들을 소거해나가는 데 결정적인 역할을 한다. 반면 자율규제의 의식적 측면은 주로 지나간 대응 사례들을 반추해서 이후 더 나은 오류 소거 능력을 갖추는 데 기여한다(Bedny, Karwowski, & Bedny, 2015, 29).

휴버트 드레이퍼스(Hubert Dreyfus)는 하이데거 현상학 가운데서 세계-내-존재 개념을 차용하여 이런 몰입 상태에서의 자율규제 능력을 설명한다. 하이데거의 현상학은 인간이 현상을 체험하며 그것으로 자기의 고유한 존재를 자유롭게 구성해 나간다는 사실에 근간을 두고 현상체험의 구체적 특징들을 조명한다. 여기서 "세계-내-존재(das In-der-Welt-sein)"는 인간의 현상이해 방식을 집약적으로 표현하는 핵심 개념이다(Heidegger, 2003b, 148). 하이데거의 표현대로라면, 인간은 그 일상성 가운데서 세계와 공속적으로 존재하는 자기존재를 명시적으로 의식하지 못한 채 세계 내에 "빠져있다"(Heidegger, 2003b, 241). 인간은 일상에서 직면하는 다양한 문제들 가운데 상당부분을 확연하게 의식하지 못한 채로 그 상황에 빠져들어가서 대응한다. 이처럼 문제적 환경에 대응하는 능력이 주로 의식되지 않은 채 비명시적으로 발휘되는 것을 "이해

(das Verstehen)"의 능력이라고 한다(Heidegger, 2003b, 208). 한편, 때로는 문제상황이 평상시와 크게 다르거나 낯선 점이 있어 이해의 능력만으로는 해결할 수 없는 난관에 부딪치기도 한다. 이 때 비로소 지적 의식이 명시적으로, 그리고 주도적으로 발동하는데 이것을 "해석(die Auslegung)"이라고 한다. 여기서 해석이란 단지 텍스트의 의미를 읽어들이는 행위라는 의미가 아니라 현상 속에서 자기 실존과 연관된 의미를 지적 능력을 활용해서 의식적으로 읽어내고 파악하는 행위라는 뜻이다(Heidegger, 2003b, 218). 이해와 해석은 자로 잰 듯 구분하기는 어렵고 대개의 경우 함께 발휘되는데, 인간의 삶을 살펴보면 이해가 우세한 순간들이 대부분이지만 난관이나 어려움에 부딪치는 순간에는 해석의 능력이 그 사람의 행위를 지배하게 된다(Heidegger, 2003b, 208).

드레이퍼스는 인공지능 기술이 근본적으로 인간의 해석 행위를 구현해 놓은 것이라고 지적하면서, 지성으로 의식되는 영역 바깥에서 인간이 발휘하는 현상대응 능력, 즉 이해의 능력이 상당부분 결여되어 있다고 평가한다(Dreyfus & Dreyfus, 1986, 5). 인간도 해석의 능력을 힘써 발휘하는 상황에서는 이성적 합리성을 따지게 되고, 이런 합리성은 상당부분 수학적으로 표현이 가능하다. 따라서 인간 역시 일상적으로 해결하기 힘든 난관에 부딪칠 때는 변수를 설정하고 가측치를 확보해서 그것들 간의 적절한 관계와 조합을 검토하는 '계산'을 한다는 점은 분명하다. 다시 말해 사안의 여러 직관적 측면들을 합리적으로 고찰해보는 것이다. 하지만 이런 "숙고의 합리성(deliberative rationality)"이 문제적 현실에 적절하고 자연스럽게 대응하는 길을 마련할 수 있는 이유는 지적 의식 이상의 것들, 즉 이해의 행위가 지적 의식의 밑바탕을 먼저 마련해 놓기 때문이다(Dreyfus & Dreyfus, 1986, 36). 세계 내부에 펼쳐지는 현상의 향연 속에 몰입되어 비명시적으로 발휘하는 이해의 능력, 이것이 결여되면 해석의 능력이 현상적 실재와 괴리된 채 주체의 사고 안에 고립된다. 이는 디지털 컴퓨팅, 수학적 추상화에 의존하는 인공지능의 근본적 한계이다.

드레이퍼스는 이 한계를 보다 구체적으로 설명하기 위해 프레임 문제(frame problem)를 거론한다. 프레임 문제란 인식 주체가 어떤 문제를 해결하

기 위해 어느 범위까지의 변수들을 유념할지 결정할 때 그 범위가 적절하지 못한 경우에 발생한다. 인간의 경우 운전을 할 때 안전하고 적절한 운전을 위한 환경 요인들과 행동 요인들을 감성적 직관을 통해 시시각각 유동적으로 변화시키면서 이해의 영역 안으로 포섭한다. 반면 인공지능은 알고리즘상 고려할 수 있는 변수의 범위가 인간에 비해 고정되어 있고, 어떤 변수의 가측치에 더 큰 가중치를 부여할지 결정하는 능력도 알고리즘에 정해진 규칙들의 범위를 벗어나지 못한다(Dreyfus, 1992, 288-289). 현재 개발되어 활용되고 있는 약인공지능은 근본적으로 각 상황에 맞는 변수와 가측치를 선정할 능력이 인간에 비해 부족하며, 이로 인해 인간의 입장에서는 비교적 단순한 작업이라 해도 오작동을 줄이기 위해 무수한 데이터 입력과 그에 따른 검증 과정을 거쳐야 한다. 이 막중한 부담을 줄이기 위해 인간의 일상적이고 효율적인 의사결정 방식을 모방한 간편추론 기법인 휴리스틱(heuristic)이 자주 활용되기도 한다(George, 2002, 228). 그러나 이런 경우에도 여전히 변수 및 가측치 선정의 경직성으로 인해 미리 설정되지 않은 변수에 대한 처리능력이 결여되는 사례들이 자주 목격된다. 앞서 언급한 자율주행 셔틀 사고가 이런 프레임 문제를 예시한다고 볼 수 있다. 특정 각도에서 차량이 저속으로 접근하는 상황이 해당 자율주행 시스템의 인지가능 영역, 즉 변수들의 '프레임'에 들어가 있지 않았던 것이다. 인간 운전자 역시 부주의나 인식하기 어려운 요인 때문에 교통사고를 일으키지만, 해당 자율주행 시스템이 인지의 프레임 안에 포섭하지 못한 변수는 인간 운전자라면 거의 확실히 사고의 핵심 변수로 인지할 수 있는 것이었다는 사실을 눈여겨봐야 한다. 즉 현재의 자율주행 시스템은 휴리스틱의 적용에도 불구하고 아직 프레임 문제를 극복하지 못한 한계 때문에 다양한 오작동 사고를 낼 수 있음을 확인시켜준 것이다.

인공지능의 오작동 문제와 관련해서 기술주의적 낙관론을 제시하는 이들, 즉 목표된 작업을 수행하는 연산을 고도로 정교화하여 성능을 탁월하게 높이는 것이 오작동에 대한 근본적 대안이라고 확신하는 이들은 우선 약인공지능이 가진 이 한계를 적절히 대처할 방안을 찾아야 한다. 수학적 추상화를 통해

문제상황에 적절히 대응하는 컴퓨팅 기법을 보다 면밀하게 가다듬는 것도 반드시 필요하다. 그러나 인간이 상대하거나 인간과 함께하는 물리적 현실에서 인간 수준으로, 혹은 인간보다 탁월하게 프레임 문제를 극복하는 인공지능을 개발하려면 인간의 비명시적 현상이해 방식을 분명하게 밝혀내고 이를 재현하거나 모방할 수 있어야 한다. 드레이퍼스의 견해대로라면 이는 대단한 난제가 아닐 수 없다(Dreyfus, 2007, 1138-1139).

이처럼 인공지능 오작동에 대한 기술주의적 낙관론은 현상학적인 관점으로 볼 때 근본적인 의문을 제기하게 만드는데, 현대 타자윤리의 관점으로 보면 이보다 더 중대한 문제점이 발견된다. 극단적 성능향상을 통해 윤리적 문제 발생의 가능성을 최소화하겠다는 기술주의적 접근법은 인공지능이 수학적 추상화를 수행할 때 인지에서만 아니라 윤리적으로도 인격적 관계의 현실에서 괴리되고 그로 인해 근본적으로 몰(沒)윤리적일 수밖에 없다는 사실을 간과하고 있다. 이런 몰윤리성은 인간-인공지능 상호작용에서 인간들 사이의 윤리적 관계 및 상호작용을 구현할 가능성을 크게 축소시키는 요인이 된다. 흔히들 오해하는 것과 달리 인간들 사이의 인격적 관계에서 윤리적 책임은 실천적 결단을 위한 몇몇 특정한 시점에만 부여되는 것이 아니다. 인간은 타인을 대면하는 모든 순간에 항상 그 타인의 타자성을 환대하고 존중할 윤리적 책임에 노출되어 있으며, 이런 상시적 책임이 결정적인 결단과 실천을 필요로 하는 순간에 집약적으로 의식된다는 것이 타자성과 다원성을 중시하는 현대 타자윤리의 핵심 전제이다(Indaimo, 2015, 81, 235). 이런 관점으로 본다면 몰윤리성이 지배하는 인간-인공지능 상호작용은 그 기술적 정밀함에도 불구하고 인간의 고유한 자유와 존엄성을 무시할 가능성을 상시적으로 배태하고 있다.

이런 타자윤리의 전제와 통찰을 바탕으로 인공지능 오작동과 그 기술주의적 대안을 비판적으로 반성하기 위해 본 논의에서는 오늘날 포스트모던 윤리의 근간을 이루는 실존론적 혹은 현상학적 타자윤리의 여러 지류들 가운데 특별히 에마뉘엘 레비나스(Emmanuel Levinas)의 윤리현상학이 주는 통찰들을 채택한다. 여기서 레비나스의 유대교적 타자윤리에 주목하는 이유는 다음

의 세 가지로 압축된다. 첫째, 그의 윤리현상학은 인간의 대상화하는 인식에 배태된 독단성 폭로와 지탄에 특화되어 있다(Levinas, 2018, 275-276). 둘째, 레비나스의 윤리 논의는 주체의 자율적 판단에 바탕을 둔 주체적 도덕이 아니라 타자의 존재에 대한 수동적인 체험 가운데 지켜야 할 관계 윤리에 무게중심을 둔다(Levinas, 1996, 77-78, 106). 셋째, 레비나스가 밝히는 윤리적 책임은 타자와 마주하는 현실에서 감성적으로 느끼는 숭고한 감정 혹은 감흥에 힘입어 막중한 무게감을 확보한다(Levinas, 2010, 270-273).

레비나스 윤리현상학의 이 세 가지 특징은 인공지능의 오작동에서 확인되는 수학적 추상화 기법의 폐쇄성과 독단성 문제를 검토하고 대안을 모색하기 위한 적절한 논점과 방향성을 제시하는 데 유용하다. 레비나스의 견해를 적용해 본다면 인공지능이 수행하는 수학적 추상화는 인간의 사고력 가운데 지극히 가치중립적이고 비감성적이며, 그래서 몰윤리적일 수밖에 없는 단면을 압축적으로 구현한 기술이라고 볼 수 있다. 레비나스에 의하면 인간은 기본적으로 자기 존재를 보존 유지하려는 본능(*conatus essendi*), 혹은 "나의-존재-안에-머물려는-집착(persévérer-dans-mon-être)"을 갖고 있으며, 이에 따라 기본적으로 타자의 현상을 자기 존재 안으로 흡수해버리는, 즉 타자성을 자의적으로 자기 존재의 부분으로 환원해서 소유하려는 욕망에 의해 추동되어 살아간다(Levinas, 2013, 26). 하지만 현실을 엄밀히 따져보면 인간은 자기 주위의 세상과 타인에 의존해서 살아간다. 자아의 신체는 세상으로부터 양분과 삶의 동력을 얻으며, 자아의 의식은 세상과 타인에 의해 교육받고 성장한다. 이처럼 자아의 존재는 철저하게 타자와의 관계에 의존하고 있다. 그러므로 자아는 기본적으로 타자를 자기 소유로 주장할 권리가 없다. 그렇지만 인간은 자기 존재를 보존하고 돋보이게 하려는 욕망에 따라 타자로부터 얻는 것들을 그저 자기의 것으로 여기고 "향유하며" 살아간다(Levinas, 2018, 208).

레비나스가 보기에 인간의 삶 전반을 지배하는 이 자기 존재에 대한 집착, 자기중심성이 의식을 통해 발현된 것이 대상화 혹은 추상화하는 사고방식이다(Levinas, 2010, 248-249). 칸트의 관점으로 본다면 대상화는 인간 정신의

선험적 한계 때문에 불가피하게 따라야 하는 인식의 방편이다. 하지만 칸트도 순수사변이성 단독으로는 도덕적 실천에 이르는 사고력, 즉 실천이성을 일깨울 수 없고, 자기의 자유에 대한 깊은 고민과 비판을 통해 비로소 실천이성을 발휘할 수 있다고 보았다(Kant, 2009, 68-70). 즉 도덕적 실천을 위해서는 가치중립적인 인식에만 기대어서는 안되고 도덕 자체를 위한 별도의 사유와 고민이 필요하다는 것이다. 레비나스 역시 이러한 통찰을 이어받아 타자에 대한 대상화와 이 대상화를 기반으로 수행되는 사변적 추론을 통해서는 타인에 대한 윤리적 책임의 자각과 실천이 불가능하다고 확신했다(Levinas, 2013, 91-93). 그는 대상화하는 사변에서 가장 문제시되는 특성으로 대상화에 필히 동원되는 언어적, 개념적 보편화 행태를 지목한다. 보편화란 달리 말해 타자의 고유한 현상적 개별성을 말살해서 자아의 임의적인 개념의 틀 안에 끼워맞추는 자율적이고 일방적인 사고행위를 뜻한다. 언어는 여기에 확고한 형식적 틀을 제공하여 내적 일관성이나 합리성을 갖출 수 있게 돕는다(Levinas, 2013, 189-190). 그래서 대상화하는 인식 가운데서는 타자의 존재에 대한 윤리적 존중이 처음부터 배제된다. 반면 자아와 타자의 윤리적 관계는 타자를 자율적으로 대상화해서 인식하는 지적 의식이 주도권을 갖는 상황에서 형성되지 않고 타자의 고통과 죽음에 대한 감성적 절박함이 자아의 삶 전체를 지배하는 가운데서 비로소 형성된다는 것이 레비나스의 확고한 입장이었다(Levinas, 2018, 293-295). 칸트가 감성으로 체감하는 현실의 인과성에 전혀 영향을 받지 않고 그로부터 완전히 자유로운 순수실천이성을 통해 도덕성이 발현된다고 본 반면, 레비나스는 감성을 통한 타자의 체험을 통해서만 윤리적 책임을 깨닫고 그 책임을 짊어지는 결단이 가능하다고 본 것이다. 이처럼 레비나스는 윤리의 기반을 주체 내부의 자율적 이성이 아니라 삶의 현상들을 직접적으로 그리고 수동적으로 마주하는 감성에 두었고, 이를 통해 윤리의 핵심이 폐쇄성, 내재화, 독단성을 벗어나는 데 있다고 논증하였다.

윤리의 근거를 인간의 정신 안이 아니라 감성적 현실에 지정하는 레비나스의 통찰은 인공지능 오작동과 관련된 윤리적 반성과 대응에 중요한 교훈을

준다. 인공지능의 현실인지 능력을 극대화하는 것으로 성능 문제와 윤리 문제를 동시에 해결하려는 접근법은 엄밀히 말해 칸트 인식론을 위시한 여러 근대 인식론 지류의 도덕에 관한 기본 신념을 이어받는다. 현재 인공지능 기술의 발전을 주도하는 기술기업과 엔지니어들 대다수는 근대적 이상에 부합하는 인간을 상정하면서 인간의 정신 내부에 자리잡고 있는 대상화 및 추상화 능력을 최대한 적절하게 구현하는 데 천착하고 있는 셈이다. 여기에는 정밀한 인식에 적절한 윤리가 당연하게 뒤따를 것이라는 주지주의적 신념, 다시 말해 도덕성이 감성적 조건이나 감정에 흔들리지 않는 적절한 현실인식과 냉철한 판단으로부터 발현한다는 '근대적' 인간이해가 깊게 관여되어 있다.

그러나 현대의 윤리 담론에서 냉철한 이성의 능력에 관한 과도한 신뢰는 무너진지 오래이다. 윤리에서 이성의 위치가 사라진 것은 아니라 하더라도, 인간이 윤리적 책임을 절감하고 타인에 대한 윤리를 실천하는 데서 보다 근원적이고 결정적인 원동력은 타인을 자아에 대해 일종의 초월적 존재로 맞이하는 숭고함의 감정이라는 데 여러 윤리학자들이 공감하는 추세이다. 레비나스는 이런 현대 타자윤리의 흐름에서 중심부에 위치한 인물로, 인간의 치밀한 대상화 행위에 필연적으로 수반되는 가소성(plasticity)의 문제를 지목한다(Levinas, 1989, 137). 레비나스에 의하면, 자아에 입장에서 신비일 수밖에 없는 타자의 존재가 현상으로 다가올 때, 인간은 두 가지 대응방식을 선택할 수 있다. 하나는 대상화 혹은 추상화라는 몰윤리적 대응이다. 자아의 입장에서는 인식의 한계 때문에 원형대로 파악이 불가능한 타자의 타자성, 그의 고유한 개별성을 자아가 자의적으로 해석의 틀 안에 넣고 주물처럼 "조형"해서 굳혀버리는 것이다(Levinas, 2018, 295). 이렇게 가소적으로 추상화된 타자는 자아가 임의대로 다룰 수 있는 것처럼 취급되기 때문에 결국 타자에 대해 자아가 원초적으로 지켜야 할 윤리적 책임은 소거되어 버린다. 반면, 이와 다른 대응방식도 선택할 수 있다. 바로 타자를 신비와 초월 그대로 환대하면서 그 고유성을 조심스럽게 존중하는 방식이다(Levinas, 2018, 320).

레비나스는 이런 윤리적 대응이 가능해지는 상황으로 자아가 타인의 고통

과 죽음을 절박하게 체감하는 순간을 지목한다. 그가 여러 윤리적 실천의 현장들을 목격하고서 내린 결론은 타인의 고통과 죽음이 인간 각자의 자기중심성에 배태된 폭력성, 즉 타자의 개별성을 말살하려는 습성을 여실히 드러낸다는 것이다(Levinas, 2018, 292). 유대인 현상학자로 독일 현상학을 공부했으나 제2차 세계대전 발발과 함께 프랑스 군인으로 참전해 독일군 포로가 되었으며 포로수용소에 갇혀 있는 동안 부모와 친지들 다수가 홀로코스트의 희생자가 된 레비나스의 입장에서 봤을 때 타자성을 말살하는 삶의 태도가 실제 혐오에 근간을 둔 인명 학살로 이어지는 사태는 일정한 정황적 계기만 주어진다면 얼마든지 발생할 수 있는 일이었다(Girgus, 2010, 1). 그래서 그의 타자윤리에서는 특정한 감정이 대단히 중요하게 다뤄진다. 유대인으로서 계승한 문화적 전통에 따라, 그는 타인이 압제를 받으며 고통 속에 죽어가는 때, 혹은 타인의 개별성이 무시당하며 그 존재 가치가 허물어지는 때에 자아는 유대교 토라에 기록된 십계명 중 여섯 번째 계명, "너는 살인하지 말지니라"라는 신의 계명에 압도된다고 말한다(Levinas, 2018, 294). 그리고 이 계명을 거부하지 않는 이의 안에는 타인의 존재에 깃든 존엄성, 숭고함에 대한 무한한 존경의 감정이 일어나 막중한 책임감을 가지고 윤리적 실천을 감행한다고 밝힌다(Levinas, 2018, 320). 타자의 고통과 죽음을 마주하는 감성적 체험에서 일어나는 윤리적 감정 혹은 감흥에는 인간이 초월적인 것, 아득하고 무한한 것을 마주할 때 그의 마음 전체를 휘감는 종교적 엄숙함과 무게감이 함께한다는 것이다. 여기서 종교적이라 함은 특정한 종교의 교의에 귀속된 것이 아니라, 무한한 신비의 존재 앞에 유한자인 인간이 압도되는 감성적 체험을 의미한다.

레비나스가 설명한 인간의 윤리적 각성과 실천의 근거 및 원동력은 인간-인공지능 상호작용에 수반되어야 할 인간이해에 중요한 시사점을 제공한다. 레비나스가 윤리의 필수적 근거로 지목한 타자에 대한 직접적인 감성과 그로부터 유발되는 숭고함의 감정은 인공지능의 수학적 추상화에서는 원천적으로 배제된다. 현상을 변수와 가측치로 구성된 디지털 대상으로 전환하고, 다시 이를 수리논리와 통계가 조합된 알고리즘의 연산 규칙들 안에서 연결하고 조

합하는 기제는 타자를 가소적으로 대상화하는 인간의 인식능력 가운데 가장 정밀하고 합리적인 측면을 모방, 재현한 것이다. 인공 에이전트가 이런 사고의 정밀함, 규칙에 정합하는 고도의 합리성을 갖추기 위해서는 인간이 체험하는 현상 가운데서 주어진 규칙과 무관한 다양한 감성적 측면들을 배제해야 한다. 문제는 이 과정에서 인간의 윤리적 각성과 실천의 원천이 되는 타자에 대한 감성과 그로 말미암는 감정이라는 측면 역시 원천적으로 배제된다는 점이다. 이는 애초 인공지능 컴퓨팅을 근거짓는 수학적 추상화 자체의 한계 때문이기도 하지만, 한편으로는 그 한계를 보완하려는 인공지능 개발자들의 노력이 부재한 데에도 원인이 있다. 현재 인공지능 발전을 주도하는 기술기업들과 엔지니어들 대다수는 인간이 윤리적인 삶의 자세를 갖기 위해 정밀한 인지와 합리적 추론 이외에도 반드시 별도의 추가적인 요인과 계기가 필요하다는 사실을 간과하고 있는 것으로 보인다. 그렇기 때문에 인공지능의 기능적 성능을 높이는 데 집착하기만 할 뿐이며 인간의 윤리적 행위를 모방, 재현하는 인공 에이전트의 개발에는 큰 관심을 두지 않는다. 여기에는 경제적 이윤과 산업 주도권에 집착하는 이기적인 자기중심성이 그 저변에 짙게 깔려 있다. 인공지능 기술 자체가 비윤리적인 것은 아니다. 하지만 이를 별도의 보완책 없이 인간-인공지능 상호작용의 현실에 적용하면 어느 순간 하이데거적인 이해의 결여와 레비나스적인 윤리적 감성의 배제라는 추상화의 한계에 직면해서 오작동을 유발하고 인간의 자유와 안전을 위협하는 비윤리적 기술로 돌변할 것이다. 앞서 기술한 두 건의 중대한 오작동 사례에서도 이런 윤리적 문제가 발견된다. 사고의 위기 앞에서 극도의 불안감을 표시하는 셔틀 운전자, 추락의 위기를 막기 위해 분투하는 조종사들이 죽음을 눈앞에 두고 느끼는 절망감, 이들에 대한 윤리적 감성을 인공지능 컴퓨팅을 통해 그대로 재현하는 것은 불가능하다. 하지만 이런 윤리의 감성적 계기들이 사고를 낸 인공지능의 알고리즘 안에서 애초 변수로, 연산의 대상으로 취급되지도 않고 있다는 점은 기술의 한계에 그 탓을 돌릴 것이 아니라 개발자들의 윤리에 대한 이해의 결여와 그에 따른 노력의 부재에 탓을 돌려야 한다. 이런 맥락에서 이 두 건의

사고사례는 인공지능 윤리의 부재가 곧 인공지능의 치명적인 기술적 결함으로 이어진다는 중요한 교훈을 남긴다.

그렇다면 레비나스의 타자윤리를 바탕으로 조명한 인공지능 오작동의 윤리적 문제에 대응할 만한 적절한 기술적 대안이 존재하는가? 인공지능 컴퓨팅이 디지털 컴퓨팅 방식에 의존하는 한 인공 에이전트는 수학적 추상화에 의한 몰윤리적 인지방식을 벗어날 수 없다. 그러나 근래 빠르게 발전된 심층학습 기술은 일련의 우회적 대안을 마련할 가능성을 엿볼 수 있게 해준다. 인공지능 기술이 수많은 지류들로 나뉘어 있는 만큼 결과값 산출의 독단성 방지를 위한 기술적 보완책 역시 여러 방면으로 나누어 검토할 수 있다. 그러나 여기서는 인간이 직접적인 감성적 체험과 그로부터 유발되는 원초적인 감정으로부터 현상의 유의미성을 길어낸다는 점에 착안해서 특별히 감정 인공지능(emotional artificial intelligence, EAI) 기술을 유력한 하나의 대안으로 제시하려 한다.

감정 인공지능이란 인공 에이전트가 인간이 표현하는 감정 신호들을 데이터로 수집해서 인간의 감정 상태를 학습해 분별하고, 더 나아가 인간의 감정을 존중하여 인간이 긍정적인 감정 상태를 유지할 수 있도록 인간과 상호작용하게 만드는 데 목표를 둔 기술이다. 인간의 감정상태를 분별하고 학습하는 데 활용될 수 있는 데이터로는 얼굴 표정, 체온, 심박수, 몸짓, 목소리의 억양, 사용하는 단어, 그리고 그 외 여타 신체적 표현 등이 있다(Chakraborty & Konar, 2009, 3). 이 기술과 관련해서 아루나 차크라보티(Aruna Chakraborty)와 아밋 코나르(Amit Konar)는 감정 인공지능 기술의 유효성을 가능하게 만드는 이론적 기초로서 감정의 네 가지 특징을 지목한다(Chakraborty & Konar, 2009, 5). 두 연구자에 의하면, 인간의 감정은 강렬함(intensity), 일시성(brevity), 편향성(partiality), 그리고 불안정성(instability)을 대표적인 특성으로 삼는다. 감정의 강렬함이란 여타의 정신작용, 특히 지적인 인지와 추론에 비해 상대적으로 훨씬 강렬하다는 것을 말한다. 일시성이란 특정한 감정이 인간의 심리흐름 속에서 비교적 짧은 시간만(몇 분에서 몇 시간) 지속된다는 뜻이

다. 편향성이란 감정이 통상적으로 확실하게 정해져 있는 대상, 즉 특정한 사람이나 사물, 혹은 상황에 대한 개인적인 관심에 기인한다는 것이다. 마지막으로 불안정성이란 감정이 인간 자신을 둘러싼 신체적 혹은 정신적 정황이 빠르게 변화하는 불안정한 상황에서 주로 촉발된다는 것이다(Chakraborty & Konar, 2009, 5). 감정 인공지능은 인간과 인공지능이 함께하고 있는 상황에서 인간이 느끼는 감정의 이 네 가지 특성들을 감안하여 인간이 긍정적이고 우호적인 기분을 유지하면서 인공지능을 사용할 수 있도록 인공지능을 학습시켜 개발하는 기술이다. 혹은 최소한 인간이 인공지능을 활용하고 함께 상호작용하는 데서 불안감이나 공포심, 혹은 적대적 감정이 일어나는 것을 방지하도록 하는 데에도 응용되어 활용될 수 있다.

감정 인공지능에 대한 기술기업들과 컴퓨터 엔지니어들의 관심도는 최근 크게 높아지는 중이다. 물론 이런 관심도는 주로 해당 기술의 경제적 이윤창출 가능성에 비례한다. 감정의 식별 및 인지 기술과 관련된 전 세계 시장규모는 2022년 401억 달러(한화 약 54조원)로 측정되었는데, 5년 뒤인 2026년에는 1,167억 달러(한화 156조원)로 3배 가까이 증가할 것으로 전망된다(Mordor Intelligence, 2023). 이 성장세는 인공지능이 인간의 일상 속, 특히 인간의 감정 영역에 보다 더 깊이 관여되고 있는 추세를 반영한다. 가장 초기의 감정 인공지능은 주로 정신의학 분야에서 보다 효율적이고 정확한 진단을 위한 보조도구로 개발되었으나, 현재는 이 기술이 인공지능 서비스 사용자에게 우호적인 정감을 유발하여 소비자들의 고객 충성도를 극대화하려는 목적으로 주로 활용되고 있다(Monteith et al., 2022, 204). 이렇게 상업적 이익을 증대시키기 위해 활용되는 감정 인공지능 기술은 즐거움과 위로감 등 주로 인간의 긍정적 감정을 자아내거나 지속시키는 데 주안점을 두고 개발되고 있다. 반면, 공포나 불안과 같은 인간의 부정적 감정에 대응하는 인공지능의 개발 시도는 쉽게 찾아보기 어려운데, 이는 이런 기술적 접근법이 단기적인 이윤 창출에는 별반 도움이 되지 않기 때문인 것으로 여겨진다.

레비나스 타자윤리의 관점으로는 감정 인공지능 기술이 인간의 부정적 감

정들을 식별하고 그에 부합하는 문제해결 능력을 보여주도록 설계할 때 유의미하게 활용될 수 있을 것으로 판단된다. 앞서 살펴본 바와 같이 레비나스는 기본적으로 타인의 고통과 죽음이라는 현상체험을 윤리적 각성과 결단의 결정적 계기로 바라보기 때문에 인간의 부정적 감정에 큰 의미를 두고 그것을 감지하는 데 힘을 들일 것을 권고한다. 인간의 긍정적인 감정보다 부정적 감정에 주목하는 이런 태도는 감정의 네 가지 대표적 특징에 대한 차크라보티와 코나르의 설명에도 온전하게 부합한다. 특히 감정의 불안정성이라는 측면에서 부정적 감정은 긍정적 감정보다 훨씬 더 예리한 현실인지를 가능하게 해준다. 긍정적 감정은 강렬함, 일시성, 편향성의 특성은 있으나 주위세계의 변화에 대한 불안정성에 대해서는 덜 민감한 편이다. 반면 부정적 감정은 삶의 정황에 닥친 위기와 격변을 직접적으로 반영한다. 인공지능 윤리가 인간-인공지능 상호작용의 현장에 관여되는 윤리적 정황에 대한 예민한 감성을 중시해야 한다는 관점으로 볼 때, 고통과 죽음의 위협이 닥쳐오는 상황에 대한 인간의 불안, 두려움, 절망 같은 감정의 표식들은 인간이 즉각적인 도움을 필요로 하는 급박한 처지에 놓여 있다는 것을 인지하기 위해 최우선적으로 고려되어야 할 현상적 계기들로 간주된다.

차크라보티와 코나르에 따르면(2009, 6), 인간의 감정 발흥과 관련해서 가장 중요한 요소는 “평가적 요소(evaluative component)”라 할 수 있다. 감정은 삶의 정황에 대한 평가를 수반한다. 이는 편향성 및 불안정성과 깊게 관련되어 있는데, 만일 인간이 어떤 현상에 직면해서 그 의미를 평가하려는 관심이 없다면 그 현상은 무관심과 무감정 속에 그저 지나갈 뿐이다. 즉 감정은 현상과 정황에 대한 개인적 관심과 평가의지 때문에 일어나고, 이런 관심과 의지의 정도가 클수록 감정이 일어나는 강도 또한 비례해서 높아진다. 긍정적인 감정은 삶을 둘러싼 정황이 안전하고 쾌적할 때 일어나고, 부정적인 감정은 삶의 정황이 위협적이고 불쾌할 때 강렬하게 느껴진다(Chakraborty & Konar, 2009, 6). 차크라보티와 코나르 두 사람이 설명한 감정의 이런 평가 기능은 하이데거와 드레이퍼스의 현상학에서 규명한 이해와 해석의 관점으로 볼 때

두 부류로 다시 나눌 수 있다. 하나는 이해의 능력에 의해 이루어지는 무의식적이고 즉각적인 평가이고, 다른 하나는 해석의 능력에 밑바탕을 둔 의식적이고 반성적인 평가이다. 감정이 지닌 이런 평가기능 덕분에 감정 인공지능 기술은 인공 에이전트가 직접 인지하지 못하는 인간의 삶의 정황 변화를 인간의 감정표현을 통해 간접적으로 인지할 수 있게 해준다.

인간의 감정이 지닌 이런 특성들과 그것에 의존해서 인간을 살피는 감정 인공지능의 특화된 기능을 상기 두 건의 오작동 사고사례에 접목해 생각해보면, 인간의 부정적 감정을 인지하는 감정 인공지능 기술, 그 가운데서도 특별히 인간의 실존적 이해에 힘입은 무의식적이고 즉각적인 정황의 평가를 인지하는 기술은 인공지능 오작동이 초래하는 위험으로부터 인간 사용자를 보호하는 데 유용하게 활용될 수 있을 것으로 사료된다. 특정 인공지능 시스템이 내놓은 결과값과 그로 인한 물리적 작용이 인간의 신체나 정신에 커다란 위협으로 다가올 때 해당 인공 에이전트가 인간의 표정, 음성, 몸짓, 심박수 등 감정상태에 직결된 데이터를 통해 극도로 부정적인 감정을 감지해서 문제가 되는 결과값을 즉각적으로 철회하거나 비상대응할 수 있도록 설계되어 있다면 라스 베이거스 자율주행 셔틀 사고나 MCAS에 의한 항공기 추락사고 같은 오작동 사고의 방지가 가능해질 것이다. 자율주행 차량의 경우 탑승자나 상대 운전자의 급작스러운 불안이나 공포감을 인지할 때 급히 최저속 운행으로 전환하거나 탑승자 및 여타 운전자, 보행자 안전을 위한 안전조치를 수행하도록 시스템을 설계한다면 인공 에이전트가 변수와 가측치로 포착하지 못하는 위기상황에 대한 최소한도의 대응이 가능할 것이다. MCAS의 경우 추락 위기를 맞아 조종사들이 느끼는 극도의 스트레스를 인지했을 때 인간 조종사에게 위기감을 야기하는 결과값 수행을 즉각적으로 철회하거나 인간 조종사가 작동 중지 여부를 결정할 수 있도록 하는 알고리즘 설계가 이루어진다면 인공지능 시스템의 부적절한 독단성에 의한 대형사고를 방지할 수 있을 것이다.

실제로 이런 구상과 유사한 기술적 개념들이 엔지니어들에 의해 여러 방면으로 제시되고 있다. 일례로 보잉 737 맥스 MCAS 오작동 사례를 분석한

장과 황(2020, 711-712)은 유기적 인간-컴퓨터 상호작용 설계(organic human-computer interaction design) 개념을 제안한다. 두 연구자에 따르면, 인간은 유기적인 의사결정 과정을 거치는 반면 인공지능은 비유기적인 소프트웨어 의사결정 과정을 따른다. 인공지능이 사용자를 존중하는 시스템으로 작동하려면 인공지능 소프트웨어가 인간의 유기적 의사결정 과정에 적응하는 기제를 갖춰야 한다. 유기적 의사결정 과정의 가장 중요한 특징은 "종합적(comprehensive)"이라는 점이다. 인간은 감성과 지성 전체를 활용해서 문제상황에 관여된 환경적 요인이나 인격적 요인을 포괄적으로 사태의 판단에 관여시킨다. 이 과정에서 인간은 필요한 경우 타인과 "협력적 상호작용(cooperative interaction)"을 개시할 능력, 그리고 프레임 문제에 갇히지 않고 정황에 맞게 인지하고 판단하는 "역동적 조정(dynamic adjustment)"의 능력을 갖고 있다(Zhang & Huang, 2020, 711-712). 그러므로 보잉 737 맥스의 MCAS 오작동 같은 사태를 방지하려면 인간의 유기적 의사결정 과정에서 발휘되는 이런 능력들을 보조하거나 존중하는 알고리즘 설계가 이루어져야 한다. 종합적 인지를 위해서는 여러 종류의 센서를 통해 현상과의 접점을 다각화하고, 협력적 상호작용과 역동적 조정을 위해서는 인공지능 시스템 내부에서의 자체적인 학습과 피드백 외에도 반드시 결과값 산출의 근거를 인간에게 고지해서 인간 편의 피드백을 받고 이 피드백을 우선시해서 학습을 강화시키는 방식으로 시스템 설계가 이루어져야 한다(Zhang & Huang, 2020, 713). 장과 황은 인공지능 시스템의 고질적인 프레임 문제를 극복하려면 인간의 역동적 조정 능력이 반드시 필요하고 인공지능 시스템이 이 능력에 맞춰 학습 방향을 설정해야 한다고 주장한다.

두 연구자가 제시하는 유기적 인간-컴퓨터 상호작용 설계를 위한 조건에서 눈여겨봐야 할 점은 바로 인간의 피드백이 인공지능 시스템 자체의 연산이나 학습보다 우선시되어야 한다는 점이다. 그렇다면 인간의 피드백을 우선시할 때 인간이 내린 판단의 어떤 측면을 가장 우선시해야 하는가? 두 연구자는 이를 추가적인 논의가 필요한 논제로 남겨놓았는데, 레비나스 타자윤리의 관

점으로 본다면 윤리적인 인간-인공지능 상호작용을 위해서는 인공지능의 결과값에 대한 인간의 부정적 감정표현에 즉각적으로 반응하는 감정 인공지능 기술을 적용하는 것이 효과를 보일 것으로 예견된다. 인간의 공포와 불안이라는 현상을 입력값으로 삼고, 삶의 정황 변화에 대한 평가가 담긴 인간의 감정적 피드백을 따라 학습의 방향을 조정해간다면 독단적 결과값 산출에 기인한 인공지능 오작동을 인간과 인공 에이전트가 함께 적절히 통제하고 방지하는 것이 가능해진다. 이는 인간의 신체적, 정서적 안녕에 심대한 영향을 줄 수 있는 물리적 인간-인공지능 상호작용에 반드시 요구되는 시스템 설계상 조건이라고 볼 수 있다.

마지막으로 한 가지 짚고 넘어가야 할 점은, 감정 인공지능 기술을 활용하는 오작동 방지 대안이 비록 인간의 무의식적이고 비명시적인 이해의 능력, 윤리적 감성을 활용하지만 결국에는 수학적 추상화 방식을 벗어나지 못한다는 점이다. 다시 말해서 인간에게는 강렬한 윤리적 각성의 계기가 되는 타인의 위기감과 불안, 공포에 대한 감성적 체험들 역시도 인공 에이전트에게는 수학적 추상화를 통해 인지해야 하는 디지털 대상의 하나에 불과하다. 즉, 감정 인공지능 기술을 통해 행동주의적인 방식으로 재현된 윤리적 감정은 당연하게도 인간이 체험하고 느끼는 그 자체대로의 '실재적인(real)' 윤리적 감흥이나 자각이 될 수 없다. 하지만 인공 에이전트가 인간의 의식적이고 반성적인 사고에 대응하는 것보다 인간의 감정반응을 살피고 감지하는 데 더 주력한다면 인간이 체험하는 감성적 현실과 인공지능 시스템의 추상화된 인지 사이에 발생하는 괴리의 정도를 다소간 완화시킬 수는 있다. 인공지능의 현상 인지는 추상적일 수밖에 없지만, 적어도 윤리적 감성과 직결된 인간의 실존적 이해의 영역에서 해당 시스템이 받아들이는 입력값, 즉 디지털 대상으로 전환될 현상들을 확보한다면 인공지능이 갖추지 못한 인간의 이해능력 및 윤리적 감정을 어느 정도 보완해서 재현할 수 있는 우회적 방안이 마련되는 것이다. 이는 인공지능의 자율성을 그 독단성 제어를 위해 역이용하는 방안으로서, 타자에 대한 인간의 윤리적 감성과 감정에 대한 '가상적(virtual)' 혹은 '유비적

(analogical)' 구현을 통해 인공 에이전트가 안전하고 윤리적인 문제해결 능력을 갖추도록 하는 '실효적(effective)' 대안으로 그 성격을 규정할 수 있다.

Ⅳ. 나오는 말

수학적 추상화에 의존하는 인공지능 컴퓨팅은 근본적으로 감성적 직관의 영역과 괴리되어 있고, 이로 말미암아 인간-인공지능 상호작용에서 언제든 독단성을 내보이며 오작동을 유발할 수 있다. 변수, 가측치, 관계검증 연산만으로는 인간이 겪는 다채로운 감성적 이해를 대체하거나 능가할 수 없다. 현재 수리논리와 통계연산에 의거해서 세계를 인지하고 인간과 상호작용하는 인공지능 기술의 수준으로는 이상적인 인간-인공지능 상호작용을 이뤄내는 데 여러 난관이 존재한다. 따라서 인간이 인공 에이전트와 상호작용할 때 자율성의 주도권을 전적으로 인공 에이전트에게 맡기는 것은 시기상조라 볼 수 있다. 인간의 감성적 직관이 실존적 이해와 윤리적 감정의 발원지인 한, 이 감성적 직관이 배제된 인공 에이전트의 인지와 판단이 인간과 함께하는 그 순간의 시공간적 정황에 맞는 탁월한 적실성을 확보하리라고 기대하기 어려운 것이다. 이런 적실성의 결여는 윤리의 영역에서 보다 중대한 문제로 부각된다. 레비나스의 타자윤리 관점으로 볼 때, 인공지능의 오작동 가능성은 기술 발전을 통해 경감될 수는 있을지 모르지만 근본적으로 소거될 수는 없다. 인공지능은 감성적 직관이 배제되어 있다는 근본적 한계 때문에 타자로부터의 고립과 독단성에 기인한 비윤리적 판단의 가능성을 상시적으로 내포하고 있다.

라스 베이거스 자율주행 셔틀 오작동과 보잉 737 맥스 MCAS 오작동 사례는 실제 인공지능이 인간과 상호작용하는 현실에서 직면하는 현상적 실재와의 괴리 문제를 여실하게 보여준다. 따라서 디지털 컴퓨팅에 의존하는 인공지능 시스템은 인간과 상호작용할 때, 특히 물리적인 방식으로 상호작용할 때 언제든지 인간에게 자율성의 주도권을 내맡길 태세를 갖추고서 인간의 인지

와 판단, 그리고 행동을 보조하는 '증강지능(augmented intelligence)'의 역할을 수행하는 데 주력해야 한다. 현 시점에서 완전한 자율지능의 개발은 단지 기술적으로 불가능할 뿐 아니라 현상학적으로나 윤리적인 관점으로 볼 때에도 그 가능성을 쉽게 예단하기 어렵다. 인공지능이 그 인지적 한계에 부합하는 수준의 자율성을 가지고 증강지능 역할을 수행하도록 하는 방안으로 여러 가지 기술적 대안이 제시될 수 있는데, 그 가운데 감정 인공지능이 실효적인 대안으로 입증될 가능성에 눈여겨볼 필요가 있다. 이는 레비나스 타자윤리의 통찰에 힙입은 제안으로서, 인간의 부정적 감정표현, 특히 고통과 죽음의 불안을 느낄 때 표출하는 감정표현에 반응하는 시스템 구축을 통해 인공지능 오작동을 방지하는 방안이다. 타인의 고통과 죽음이 일으키는 강렬하고 압도적인 윤리적 책임의 감흥을 강력한 윤리적 실천 동기로 지목하는 레비나스의 인간학적 사유에 따라, 인공 에이전트가 인간의 부정적 감정을 감지할 때 그 결과값의 적실성을 재고하고 인간 편에 판단의 주도권을 내어주는 윤리적 인공지능 시스템을 개발, 적용한다면 인간에게 부정적인 영향이나 효력을 발생시키는 인공지능의 독단성을 실효적으로 통제할 수 있을 것으로 사료된다. 이는 인간만이 누리는 종교적-윤리적 감성과 감정의 유비적 재현이라고 볼 수 있다.

【주제어】 인공지능 오작동, 수학적 추상화, 레비나스, 윤리적 감성과 감정, 감정 인공지능

[참고문헌]

이종권 (2005). 칸트의 직관 개념과 수학적 직관주의. 칸트연구, 15, 191-222.

Descartes, R. (1982). Discours de la Méthode, Regulae ad Directionem Ingenii. 이현복 옮김 (2013). 방법서설, 정신지도를 위한 규칙들. 서울: 문예출판사.

Heidegger, M. (1983). Die Grundbegriffe der Metaphysik: Welt-Endlichkeit-Einsamkeit. 이기상, 강태성 옮김 (2001). 형이상학의 근본개념들: 세계-유한성-고독. 서울: 까치글방.

_____. (1991). Kant und das Problem der Metaphysik. 이선일 옮김 (2003a). 칸트와 형이상학의 문제. 파주: 한길사.

_____. (1993). Sein und Zeit. 이기상 옮김 (2003b). 존재와 시간. 서울: 까치글방.

Husserl, E. (1976). Ideen zu Einer Reinen Phänomenologie und Phänomenologischen Philosophie, Bd. 1. 이종훈 옮김 (2009). 순수현상학과 현상학적 철학의 이념들 1: 순수현상학의 일반적 입문. 파주: 한길사.

Kant, I. (1998). Kritik der reinen Vernunft. 백종현 옮김 (2006). 순수이성비판 1. 서울: 아카넷.

_____. (2003). Kritik der praktischen Vernunft. 백종현 옮김 (2009). 실천이성비판. 서울: 아카넷.

Levinas, E. (1979). Le temps et l'autre. 강영안 옮김 (1996). 시간과 타자. 서울: 문예출판사.

_____. (1974). Autrement qu'être ou au-delà de l'essence. 김연숙, 박한표 옮김 (2010). 존재와 다르게, 본질의 저편. 고양: 인간사랑.

_____. (1993). Dieu, la Mort et le Temps. 김도형, 문성원, 손영창 옮김 (2013). 신, 죽음 그리고 시간. 서울; 그린비출판사

_____. (1988). Totalité et infini: essai sur l'extériorité. 김도형, 문성원, 손영창 옮김 (2018). 전체성과 무한: 외재성에 대한 에세이. 서울: 그린비출판사.

Bedny, G. Z., Karwowski, W., & Bedny, I. (2015). Applying Systematic-Structural Activity Theory to Design of Human-Computer Interaction Systems. Boca Raton: CRC Press.

Chakraborty, A. & Konar, A. (2009). Emotional Intelligence: A Cybernetic Approach. Berlin: Springer.

Chanda, S. S. & Banerjee, D. N. (2022). Omission and commission errors underlying AI failures. AI & Society 2022 (Online), 1-24.

Dreyfus, H. L. & Dreyfus, S. E. (1986). Mind over Machine: The Power of Human

Intuituion and Expertise in the Era of the Computer. New York: The Free Press.
Dreyfus, H. L. (1992). What Computers Still Can't Do: A Critique of Artificial Reason. Cambridge: MIT Press.
_____. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. Artificial Intelligence, 171, 1137-1160.
Floridi, L. (2011). The Philosophy of Information. Oxford: Oxford University Press.
George, S. E. (2002). Heuristics in medical data mining. In Sarker, Abbass, Newton eds. Heuristics and Optimization for Knowledge Discovery. Hershey: Idea Group Publishing, 226-240.
Gill, K. S. (2008). Rethinking the interaction architecture. In Gill ed. Cognitioin, Communication and Interaction: Transdisciplinary Perspectives on Interactive Technology. London: Springer, 213-234.
Girgus, S. B. (2010). Levinas and the Cinema of Redemption: Time, Ethics, and the Feminine. New York: Columbia University Press.
Henderson, A., Harbour, S., & Cohen, K. (2022). Toward airworthiness certification for artificial intelligence (AI) in aerospace systems. 2022 IEEE/AIAA 41st Digital Avionics Systems Conference (DASC), 1-10.
Heyting, A. (1983). The intuitionist foundations of mathematics. In Benacerraf, Putnam eds. Philosophy of Mathematics: Selected Readings. Cambridge: Cambridge University Press, 52-61.
Hui, Y. (2016). On the Existence of Digital Objects. Minneapolis: University of Minnesota Press.
Indaimo, J. A. (2015). The Self, Ethics and Human Rights: Lacan, Levinas & Alterity. New York: Routledge.
Levinas, E. (1989). Reality and its shadow. In Hand ed. The Levinas Reader. Oxford: Basil Blackwell, 129-143.
Mongan, J. & Kohli, M. (2020). Artificial intelligence and human life: five lessons for radiology from the 737 MAX disasters. Radiology: Artificial Intelligence, 2(2), 1-3.
Monteith, S., Glenn, T., Geddes, J., Whybrow, P. C., & Bauer, M. (2022). Commercial use of emotion artificial intelligence (AI): implications for psychiatry. Current Psychiatry Reports 24, 203-211.
Mordor Intelligence (2023). Emotion detection and recognition market size & share

analysis: growth trends & forecasts (2023 – 2028). Mordor Intelligence. Available: https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market

Munn, L. (2022). The uselessness of AI ethics. AI and Ethics, 3, 869-877.

Naor, M., Adler,, N., Pinto, G. D., & Dumanis, A. (2020). Psychological safety in aviation new product development teams: case study of 737 MAX airplane. Sustainability, 12(21), 1-15.

National Transportation Safety Board (2017). Low-speed collision between truck-tractor and autonomous shuttle, Las Vegas, Nevada, November 8, 2017. NTSB. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAB1906.pdf

Picard, A., Mualla, Y., Gechter, F., & Galland, S. (2023). Human-computer interaction and explainability: intersection and terminology. In Longe ed. Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023 Proceedings, Part II. Cham: Springer Nature, 214-236.

Rose, K. (2022). The COVID-19 Pandemic: A Global High-Tech Challenge at the Interface of Science, Politics, and Illusions. London: Elsevier.

Russell, S., Moskowitz, I. S., & Raglin, A. (2017). Human information interaction, artificial intelligence, and errors. In Lawless, Mittu, Sofge, Russell eds. Autonomy and Artificial Intelligence: A Threat or Savior?. Cham: Springer Nature.

Saitta, L. & Zucker, J. (2013). Abstraction in Artificial Intelligence and Complex Systems. New York: Springer Science+Business Medias.

Schildt, H. (2020). The Data Imperative: How Digitalization is Reshaping Management, Organizing, and Work. Oxford: Oxford University Press.

Schmidt, C. T. A. (2007). An Empirically Terminological Point of View on Agentism in the Artificial. In Gelbukh, Morales eds. MICAI 2007: Advances in Artificial Intelligence: 6th Mexican International Conference on Artificial Intelligence, Aguascalientes, Mexico, November 4-10, 2007 Proceedings, 348-358.

Washizaki, H. (2022). Towards software co-engineering by AI and developers. In Virvou, Tsihrintzis, Bourbakis, Jain eds. Handbook on Artificial Intelligence-Empowered Applied Software Engineering, vol. 1; Novel Methodologies to

Engineering Smart Software Systems. Cham: Springer Nature, 39-54.
Zhang, H. & Huang, B. (2020). An organic design for human-computer interaction. In Long, Dhillon eds. Man-Machine-Environment System Engineering: Proceedings of the 20th International Conference on MMESE. Singapore: Springer Nature, 707-714.

[국문초록]

수학적 추상화에 의존하는 인공지능 컴퓨팅은 근본적으로 감성적 직관의 영역과 괴리되어 있고, 이로 말미암아 인간-인공지능 상호작용에서 언제든 독단성을 내보이며 오작동을 유발할 수 있다. 현대 현상학 관점으로 볼 때, 인공지능의 오작동 가능성은 기술 발전을 통해 경감될 수는 있을지 모르지만 근본적으로 소거될 수는 없다. 라스 베이거스 자율주행 셔틀 오작동과 보잉 737 맥스 MCAS 오작동 사례는 실제 인공지능이 인간과 상호작용하는 현실에서 직면하는 인지와 현상적 실재의 괴리 문제를 여실하게 보여준다. 인공지능이 그 인지적 한계에 부합하는 수준의 자율성을 가지고 증강지능 역할을 수행하도록 하는 방안으로 여러 가지 기술적 대안이 제시될 수 있는데, 그 가운데 감정 인공지능이 실효적인 대안으로 입증될 가능성에 눈여겨볼 필요가 있다. 이는 레비나스 타자윤리의 통찰에 힙입은 제안으로서, 인간의 부정적 감정표현, 특히 고통과 죽음의 불안을 느낄 때 표출하는 감정표현에 반응하는 시스템 구축을 통해 인공지능 오작동을 방지하는 방안이다. 인공 에이전트가 인간의 부정적 감정을 감지할 때 그 결과값의 적실성을 재고하고 인간 편에 판단의 주도권을 내어주는 윤리적 인공지능 시스템을 개발, 적용한다면 인간에게 부정적인 영향이나 효력을 발생시키는 인공지능의 독단성을 실효적으로 통제할 수 있을 것으로 사료된다.

[Abstract]

Artificial Intelligence Errors Caused by Mathematical Abstraction
- An Exploration of Technical Alternatives Based on Levinasian Consideration of Ethical Sensibility and Emotion

Park, Wook Joo (Yonsei University)

Artificial intelligence computing, which relies on mathematical abstraction, is fundamentally detached from the realm of emotional intuition. This detachment can lead to AI behaving arbitrarily and causing malfunctions in human-AI interactions. From the perspective of contemporary phenomenology, the possibility of AI malfunctions can be mitigated by technological progress, but cannot be fundamentally eliminated. The malfunctions of a self-driving shuttle in Las Vegas and the Boeing 737 MAX MCAS clearly show the problem of the gap between the phenomenal reality and AI's representations in human-AI interactions. There are a number of technical alternatives that can be proposed for AI to play an augmented intelligence role with a level of autonomy that is consistent with its cognitive limitations. Among them, it is necessary to pay attention to the possibility that emotional AI will be proven to be an effective alternative. According to Levinas's philosophical thought, developing an ethical AI system that allows artificial agents to reconsider the validity of their results when they detect negative human emotions and give humans the lead in judgment would enable humans to effectively control the arbitrariness of AI that produces negative effects or consequences for humans.

【Keywords】 artificial intelligence errors, mathematical abstraction, Levinas, ethical sensibility and emotion, emotional AI

논문투고일: 2023년 10월 30일 / 논문심사일: 2023년 12월 14일 / 게재확정일: 2023년 12월 18일

【저자연락처】 pwjjiml@yonsei.ac.kr